Regression Error Characteristic Surfaces

Luís Torgo
LIACC/FEP, University of Porto
Rua de Ceuta, 118, 6.
4050-190 Porto, Portugal
ltorgo@liacc.up.pt

ABSTRACT

This paper presents a generalization of Regression Error Characteristic (REC) curves. REC curves describe the cumulative distribution function of the prediction error of models and can be seen as a generalization of ROC curves to regression problems. REC curves provide useful information for analyzing the performance of models, particularly when compared to error statistics like for instance the Mean Squared Error. In this paper we present Regression Error Characteristic (REC) surfaces that introduce a further degree of detail by plotting the cumulative distribution function of the errors across the distribution of the target variable, i.e. the joint cumulative distribution function of the errors and the target variable. This provides a more detailed analysis of the performance of models when compared to REC curves. This extra detail is particularly relevant in applications with non-uniform error costs, where it is important to study the performance of models for specific ranges of the target variable. In this paper we present the notion of REC surfaces, describe how to use them to compare the performance of models, and illustrate their use with an important practical class of applications: the prediction of rare extreme values.

Categories and Subject Descriptors: D.2.8 [Software Engineering]: Metrics - performance measures

General Terms: Measurement, Performance

Keywords: Model comparisons, evaluation metrics, regression problems

1. INTRODUCTION

This paper addresses the issue of comparing the predictive performance of models in regression applications where the cost of errors varies across the range of the continuous target variable. We present the notion of Regression Error Characteristic (REC) surfaces, that are a generalization of REC curves [1]. REC curves draw the cumulative distribution of the errors of models. This allows comparing different models across their range of errors and thus provides more information than a statistic of errors, like for instance the Mean Squared Error (MSE). Any point in a REC curve provides an estimate of the probability of the error being less or equal to the respective X-axis value, i.e. $P(\varepsilon \leq \epsilon_i)$. These estimates provide useful information on the errors of a model. However, REC curves do not show how particular errors are distributed across the range of the true values of the target variable. This means that we can have two different models with exactly the same estimated probability of the prediction error being less than e_i, but the observed errors leading to this estimate having occurred for different true values of the target. This is not a problem in applications where the cost of any prediction error is uniform, i.e. independent of the true values of the target. However, there are applications where this is clearly not the case and we may thus require further information on how the errors are distributed across the target variable range. This is the main goal of the work presented in this paper. We add a new dimension to REC curves leading to REC surfaces. The new dimension is the range of the target variable, and thus we are able to compare the distribution of certain errors across the range of the target variable.

2. REGRESSION ERROR CHARACTERISTIC (REC) CURVES

Bi and Bennet [1] have presented REC curves. These curves play a role similar to ROC curves (e.g. [2, 3, 4]) in classification tasks, but for regression problems. They provide a graphical description of the cumulative distribution function of the error of a model, i.e. $D(\epsilon) = P(\varepsilon \leq \epsilon)$. The authors describe a simple algorithm for plotting these curves based on estimating the probabilities using the observed frequencies of the errors.

REC curves provide a better description of a model predictive performance when compared to prediction error statistics because they illustrate its performance across the range of possible errors. It is thus possible to extract more information by comparing the REC curves of two alternative models than with the two respective error statistics. Moreover, the interpretation of REC curves is quite appealing to non-experts and it is possible to obtain the same quantitative information given by prediction error statistics by calculating the Area Over the Curve (AOC), which Bi and Bennet [1] have proved to be a biased estimate of the expected error of a model.

Figure 1 shows an example of the REC curves of three models. This example shows a model (model A) clearly

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.

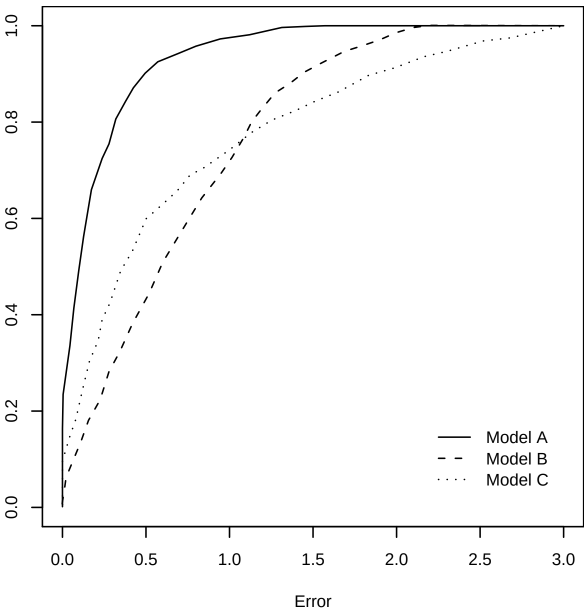

Figure 1: An example of the REC curves of three models.

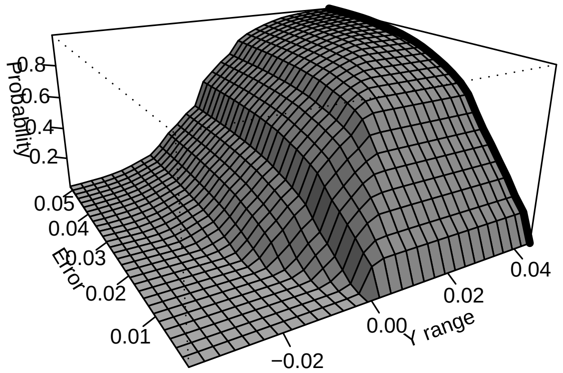

Figure 2: An example of the REC surface.

dominating the others over all range of possible errors. On the contrary models B and C have performances that are harder to compare. For smaller errors model C dominates model B, but as we move towards larger errors we see model B overcoming model C. The decision on which of these two is preferable may be domain dependent, provided their area over the curve (i.e. expected error) is similar.

In spite of the above mentioned advantages there are some specific domain requirements that are difficult to check using REC curves. These have to do with domains where the cost of errors is non-uniform, i.e. where the importance of an error with an amplitude of say 1.2, can be different depending on the true target variable value. For this type of applications, it may be important to inspect the distribution of the errors across the distribution of the target variable. In effect, it is possible to have two different models with exactly the same REC curves but still one being preferable to the other just because smaller errors occur for target values that are more relevant (e.g. have higher cost) for the application being studied. Distinguishing these two models and checking that one behaves more favorably than the other is not possible with the information provided by REC curves. This is the objective of our work.

3. REGRESSION ERROR CHARACTERISTIC (REC) SURFACES

In order to analyze how certain type of errors are distributed across the range of values of the target we propose extending the idea of REC curves to REC surfaces. The idea is to study the joint cumulative distribution function,

$$D(\epsilon, y) = P(\varepsilon < \epsilon, Y < y) = \int_0^{\epsilon} \int_{-\infty}^{y} p(\epsilon, y)\, d\epsilon\, dy \qquad (1)$$

REC curves are a particular case of these surfaces, namely a REC curve is equivalent to $D(\epsilon, \infty)$. This means that each point of a REC curve is “unfolded” into a cumulative distribution function in a REC surface, showing how the errors corresponding to that point are distributed across the range of the target variable.

Figure 2 shows an example of a REC surface obtained along the lines described above. As mentioned before the horizontal axes of the plot are the errors (as in REC curves) and the target variable (Y) range. Each point in the surface represents an estimated probability $\hat{P}(\varepsilon < \epsilon_i, Y < y_j)$. The strong bold line at the end of the “Y-range” axis is the REC curve of this model, i.e. it corresponds to $D(\epsilon, \infty)$.

REC surfaces uncover information that is not present in REC curves. Namely, there are two types of analysis that can be very relevant for certain applications, which are possible by zooming in on particular areas of REC surfaces. One is to study how a certain range of errors, $\epsilon_1 < \varepsilon < \epsilon_2$, is distributed along the domain of the target variable, i.e. on which Y values are these errors more frequent. This corresponds to a slice of the surface parallel to the Y-range axis, i.e.

$$P(\epsilon_1 < \varepsilon < \epsilon_2, Y < y) = \int_{\epsilon_1}^{\epsilon_2} \int_{-\infty}^{y} p(\epsilon, y)\, d\epsilon\, dy \qquad (2)$$

For instance, suppose we are particularly interested in checking where smaller errors are being obtained for the model whose errors are shown in Figure 2. We could zoom in the slice of the surface corresponding to $P(0 < \varepsilon < 0.015, Y < y)$, which is shown on Figure 3. This tells us that smaller errors are mostly concentrated on values of Y around zero[1], given that the surface is completely flat on the extremes of the Y distribution, though there are samples with true target value far from zero as it can be confirmed from checking the complete surface in Figure 2.

The second interesting question that can be analyzed us-

[1] Please notice the different scale of the Y variable on this figure when compared to Figure 2.

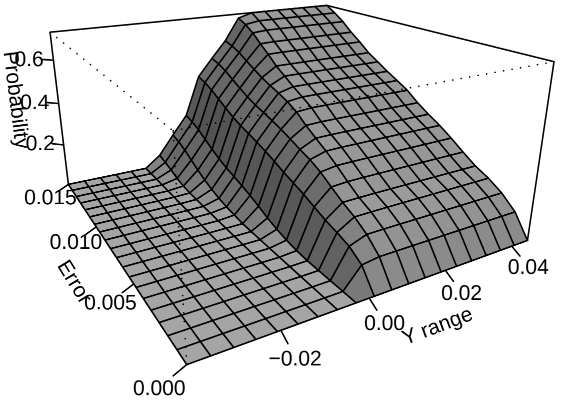

Figure 3: An example of the study of a particular range of errors.

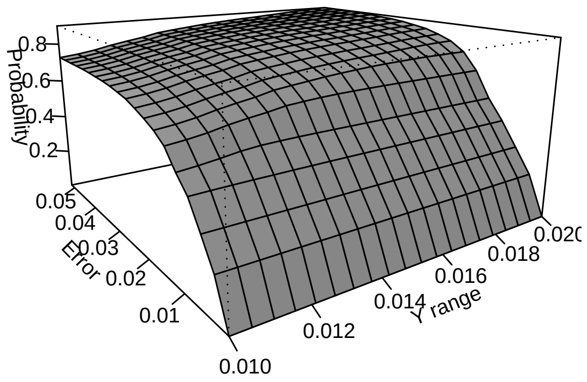

Figure 4: An example of the study of a particular range of Y values.

ing REC surfaces is to check what is the distribution of the errors for a certain range of the target variable, $y_1 < Y < y_2$, i.e. which type of errors are made by the model for this range. Again this corresponds to a surface slice this time parallel to the error axis and defined by,

$$P(\varepsilon < \epsilon, y_1 < Y < y_2) = \int_0^{\epsilon} \int_{y_1}^{y_2} p(\epsilon, y)\, d\epsilon\, dy \qquad (3)$$

We exemplify this type of slices using again the data from Figure 2. Suppose that in this particular application it is critical to have good predictions for Y values between 0.01 and 0.02. We could inspect this model performance by plotting $P(\varepsilon < \epsilon, 0.01 < Y < 0.02)$, which leads to Figure 4. This figure shows very few changes, which means that there are few testing cases in these conditions (the REC curve at $Y = 0.01$ is very similar to the REC curve at $Y = 0.02$). Still, we may notice that most surface variations occur for large errors because for smaller errors the cumulative Y distribution stays almost unchanged. This means that the few Y cases that fall on this range lead to large errors.

In order to plot a REC surface we use an algorithm with some similarities to the one presented by Bi and Bennet [1] for REC curves. Still, given that we are producing a surface we have to choose a grid of points on the Error and Y axes and then estimate the corresponding probability (Equation (1)), using the observed errors. We can complete the surface using these evaluation points in the grid with some sort of interpolation algorithm (the figures in the paper were obtained using a linear interpolation algorithm). An implementation of these ideas in the R language[2] [5] can be obtained at the following web site:
`http://www.liacc.up.pt/~ltorgo/KDD05/RECsurf.R`

4. USING REC SURFACES FOR MODEL COMPARISONS

This section describes some example uses of the information provided by REC surfaces, namely in comparing the performance of models in applications with non-uniform costs that are the main justification for using these surfaces.

REC surfaces as described in Section 3 do not bring much advantages when compared to scatter plots of predicted versus actual values of the target variable. Their advantage over these plots only arises when comparing multiple models. The same observation was made by Bi and Bennet [1] regarding REC curves. Comparing the performance of several models by superimposing their scatter plots is confusing, while plotting several REC curves on a graph (like seen on Figure 1) allows easy and clear comparisons. However, plotting several surfaces on the same graph leads to an over-cluttered image that is hard to analyze, unless different colours are used for each model surface. In order to overcome this difficulty of REC surfaces we propose to use iso-lines of the surface.

4.1 Partial REC curves

A partial REC curve is a particular case of a REC curve where we limit the analysis of the error distribution to a certain Y range. They are a two dimensional representation of the part of the REC surface defined by $P(\varepsilon < \epsilon, y_1 < Y < y_2)$. They serve the purpose of enabling an easy comparison of different surface slices that are of special interest to us. They only make sense for ranges were the cost of the errors are similar, otherwise they would be subject to the same drawbacks as standard REC curves.

Given that, $P(\varepsilon < \epsilon, y_1 < Y < y_2) = P(\varepsilon < \epsilon, Y < y_2) - P(\varepsilon < \epsilon, Y < y_1)$ we can plot a curve using the estimates of these two probabilities. We call these curves partial REC curves.

Partial REC curves can be obtained in the same way as

[2] http://www.r-project.org

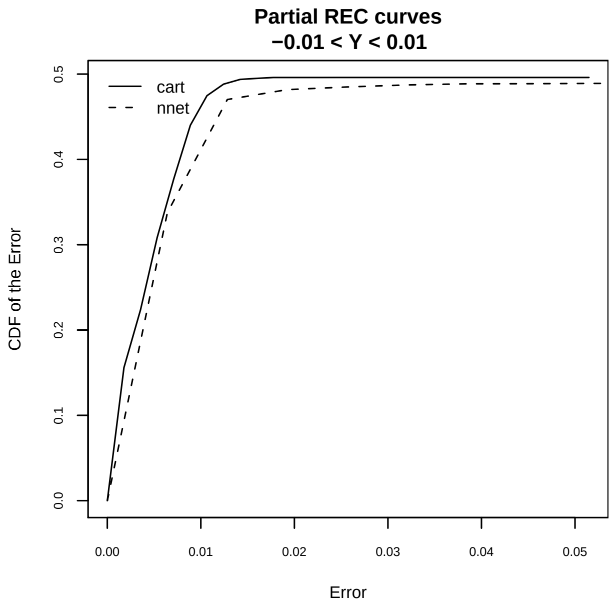

Figure 5: The partial REC curves of two models.

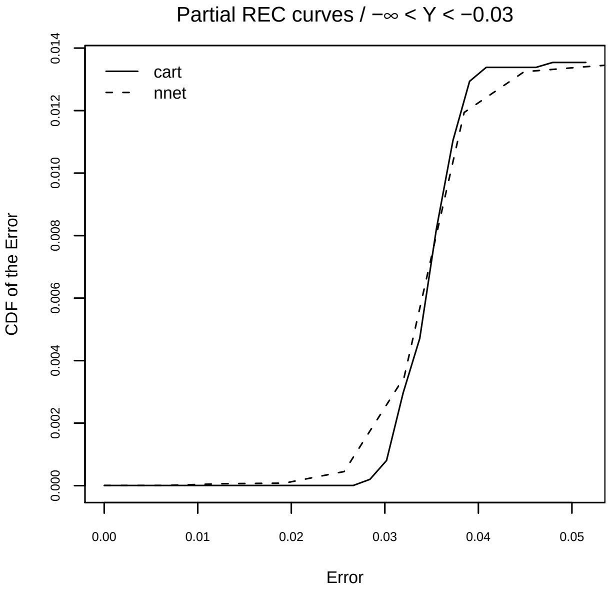

Figure 6: The partial REC curves of the same models on another Y range.

REC curves by simply estimating the necessary probabilities over the test cases in the Y range of interest. Alternatively, they may be obtained directly from the REC surface by subtracting the values of the two Y iso-lines corresponding to y_2 and y_1.

If we want to compare the performance of several models over the same Y range, we plot their respective partial REC curves and then proceed with the same analysis as with any REC curve [1].

Figure 5 shows an example of comparing the performance of two models on the same data used in previous graphs over the range $-0.01 < Y < 0.01$. On this figure we can see clearly one model dominating the other over this range of the target variable. It is interesting to observe that if we consider the range $-\infty < Y < -0.03$ instead (c.f. Figure 6), the advantage of the "cart" model is not so clear and actually the "nnet" makes smaller errors more frequently on this other range. This type of observations, which are not possible with standard REC curves, may be of crucial importance for applications with non-uniform costs, as they may help in making a more informed model selection.

4.2 Partial Y CDF's

Similarly to the analysis described in Section 4.1, we can consider the problem of comparing how a certain range of errors is distributed across the domain of the target variable. This can be used, for instance, to check whether they are particularly concentrated on a certain part of the Y domain.

This can be accomplished through a partial CDF of the Y variable for the test cases where the prediction error falls within our range of interest, i.e. $P(\epsilon_1 < \varepsilon < \epsilon_2, Y < y)$. Plotting these curves involves estimating the CDF using the subset of the test cases for which the models achieve a prediction error within our target interval.

We illustrate this use of REC surfaces in Figure 7. This figure shows the distribution of the Y variable for the cases where two models achieve a small error ($0 < \varepsilon < 0.005$). We can see that the "cart" model achieves more frequently such small errors, which is reflected in a higher final value of the CDF curve. This value corresponds to $P(0 < \varepsilon < 0.005, Y < \infty)$, which means that in a standard REC curve "cart" would dominate "nnet" for $\varepsilon = 0.005$. Nevertheless, we can also observe that these errors are achieved on different parts of the Y domain. Namely, for "cart" these errors are mostly achieved for values of the target variable near zero, because its CDF is completely flat on the extremes of the Y range. On the contrary, the small errors of the "nnet" model are slightly more widespread. In Section 5 we will describe applications where this performance of the "nnet" model would be considered preferable, which would be hard to check in a standard REC curve.

5. RARE EXTREME VALUES PREDICTION

This section describes a particular class of applications where the analysis presented in Section 4 is very useful. The main goal of this class of problems is to obtain predictive models that are able to accurately predict rare extreme values of a continuous variable. These problems are quite difficult because they involve focusing on cases that are rare (they share several features with unbalanced classification problems (e.g. [4]) and may even require specific modeling techniques (e.g. [7]). The typical distribution of the target variable of these problems is similar to a normal distribution but with extremely long tails. Examples of this class of problems include the prediction of catastrophic phenomena (e.g. harmful algae blooms in rivers [6]), or the prediction of unusual high (low) returns in financial quote data. We will use this latter problem to illustrate some applications of REC surfaces analysis.

Financial trading based on the prediction of the future returns of financial assets involves obtaining models that are able to anticipate large movements of prices. These high (or low) returns are rare, but they are the only that are interesting to traders because they allow for highly rewarding

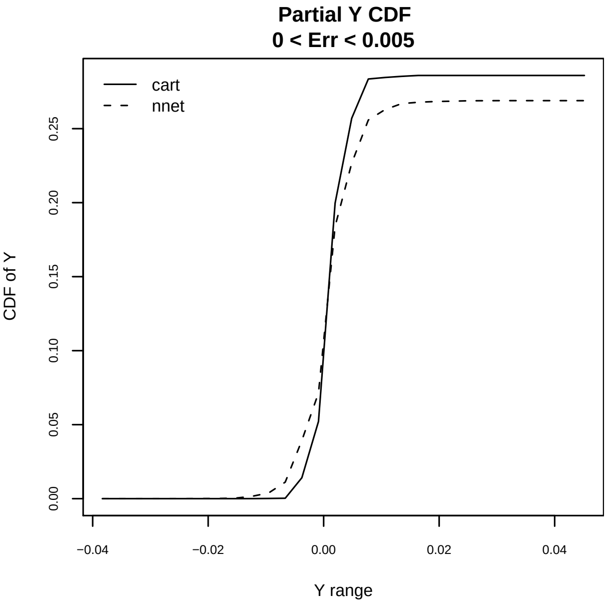

Figure 7: The partial CDF of the Y variable of two models.

trades if correctly anticipated. The typical distribution of the returns of a financial asset is centered around zero with rare low or high returns. Prediction errors on these extreme values have a much larger cost (benefit) when compared to the "normal" small returns.

Given the nature of this problem it is important to have tools that are able to evaluate the advantages of different models in terms of what is really relevant for this application. Namely, it is important to compare the performance of different models on extremely low (high) returns observed in the test cases. It is also very important to understand where large errors are being made by the alternative models we may be considering. For instance, suppose a model predicts that an unusual high return is approaching. This may lead a trader to invest a huge amount in the asset. If the true return is small, this could mean a very high cost for the trader. On the contrary, if the model predicts a small return but in effect the market shows an unusual movement then the cost is not so high, in effect it is a lost opportunity to make money but it is not a loss of money. Both these situations can be studied using the tools described in Sections 4.1 and 4.2. For instance, we can use partial REC curves to study the behavior of alternative models on extremely high and low values of the target variable (the returns of the assets). We can also use partial CDF's of the Y variable to study where a certain range of errors are occurring. Let us see some examples of this analysis.

5.1 Performance on extreme values

In this section we present some graphs illustrating example comparisons of several models on the task of predicting the daily returns of "IBM" stock prices over 10 years. We focus on the question of which model performs better on the cases that count, i.e. extreme high and low returns.

Figure 8 shows the partial REC curves of three models on extreme negative (less than 2%) and positive (higher than 2%) returns.

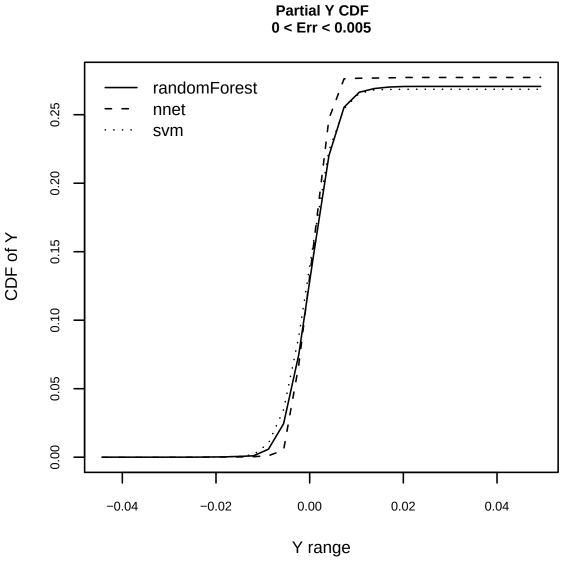

Figure 9: The partial CDF of the Y variable of the three models for small errors.

The results of the three models are quite similar. Still, we can observe that both the "randomForest" and the "svm" models are slightly superior to the "nnet" model because they have a higher percentage of smaller prediction errors for this range of returns. This means that their predictions are more accurate on the test cases that really count for this application.

5.2 Characterization of small errors

We now consider the analysis of the localization of small errors, i.e. in which parts of the target variable range are they more frequent. We again consider the prediction of the daily returns of "IBM" stocks over a period of 10 years. Plotting the partial Y CDF of the test cases for which the error was smaller than a given threshold we are able to answer such questions. Figure 9 shows such graph. We can see that the "nnet" model achieves more frequently these small errors. However, these good predictions are mostly concentrated on values of Y around zero, which are irrelevant from the perspective of a trader. In effect, outside a very small band around zero the CDF corresponding to the "nnet" model is completely horizontal. On the contrary, both "randomForest" and "svm" are able to achieve small errors on larger (smaller) returns, which is more interesting for the investor.

This is a good example of the sort of information that we cannot get from standard REC curves as described by Bi and Bennet [1]. In effect, using such curves we would observe that "nnet" would have a dominance over the other two models in small errors. However, we would not get the information that, although less frequently, the other two models achieve this type of errors on more "useful" test cases and thus we should prefer them over the "nnet" model.

6. CONCLUSIONS

In this paper we have presented a generalization of REC

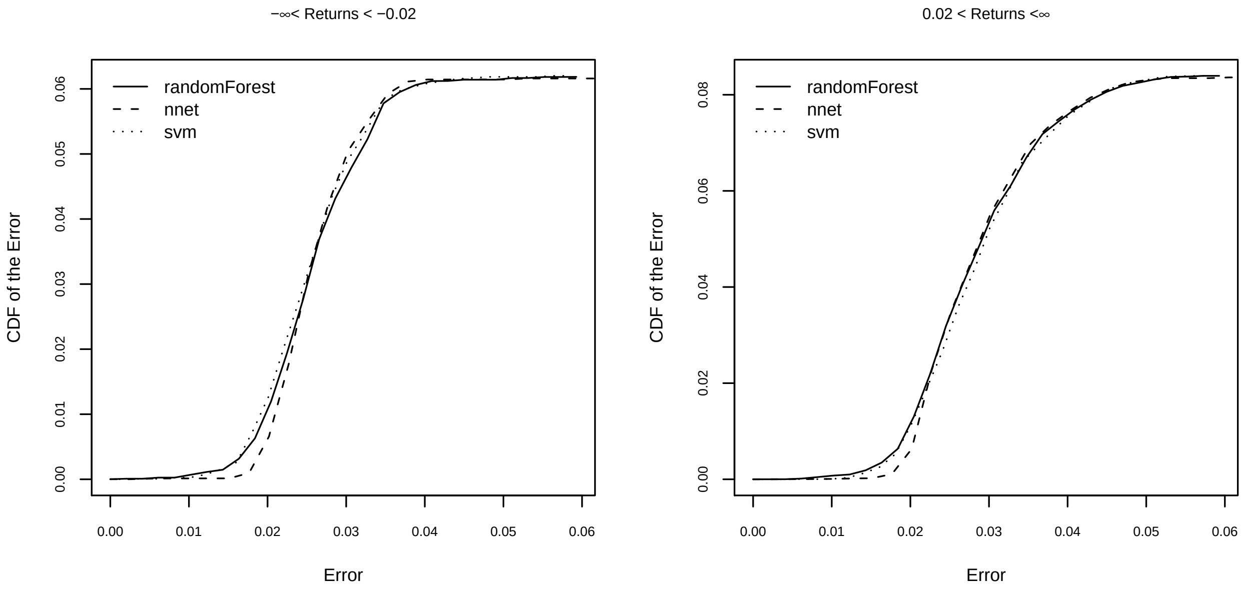

Figure 8: The partial REC curves of three models on extreme daily returns of "IBM" stock prices.

curves with the main goal of providing means to facilitate the analysis of the predictive performance of regression models on applications with non-uniform error costs. On this type of applications it is of key importance to understand where, in the range of the target variable, certain errors occur. The methodology we have presented provides this sort of information. By analyzing the REC surface of a model we are able to understand and answer several important questions concerning the performance of the model,

- On which type of values of the target variable are certain errors (e.g. small errors) more frequent?
- Which type of errors does a model make for a certain range of the target variable that is particularly important in our application?

These questions could not be answered by standard error statistics or even by looking at the REC curve of a model.

In spite of their advantages, REC surfaces are hard to analyze when comparing different models. In effect, plotting several surfaces on the same graph is messy and hard to understand. In order to overcome this difficulty we have presented partial REC curves and partial Y CDF's that are bi-dimensional representations of parts of the REC surface of a model. These curves summarize the surface of a model for a particular region of interest to us. Moreover, they allow easy comparison of the performance of several models using the same sort of analysis described by Bi and Bennet [1].

We have illustrated the concepts presented in the paper through a concrete class of applications: the prediction of rare extreme values. This is an important class of problems where it is of key importance to analyze the performance of models on particular types of values of the target variable. We have shown how to use partial REC curves and partial Y CDF's with this purpose.

We hope that the analysis made possible by the use of REC surfaces can provide insights on new forms of developing models that are tunned for a particular class of applications. In effect, we plan to explore the possibility of feeding back the information resulting from analyzing partial REC curves and partial Y CDF's into the model construction phase, thus developing models that are more useful for particular applications.

7. ACKNOWLEDGMENTS

This work was partially supported by FCT project MODAL (POSI/SRI/40949/2001) co-financed by POSI and by the European fund FEDER, and by the Faculty of Economics of the University of Porto.

The author would also like to thank Foster Provost for useful comments and suggestions on early stages of this work during the author's stay at Stern School of Business, NYU.

8. REFERENCES

[1] J. Bi and K. P. Bennett. Regression error characteristic curves. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

[2] J. P. Egan. *Signal Detection Theory and ROC Analysis.* Series in Cognition and Perception. Academic Press, 1975.

[3] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, Hewlett Packard, 2003.

[4] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proc. 15th International Conf. on Machine Learning*, pages 445–453. Morgan Kaufmann, San Francisco, CA, 1998.

[5] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-07-0.

[6] R. Ribeiro and L. Torgo. Predicting harmful algae blooms. In F. M. Pires and S. Abreu, editors, *Proceedings of Portuguese AI Conference (EPIA'03)*, number 2902 in LNAI, pages 308–312. Springer, 2003.

[7] L. Torgo and R. Ribeiro. Predicting outliers. In N. Lavrac, D. Gamberger, L. Todorovski, and H. Blockeel, editors, *Proceedings of Principles of Data Mining and Knowledge Discovery (PKDD'03)*, number 2838 in LNAI, pages 447–458. Springer, 2003.