

# Censored Time Trees™ for Predicting PSA Recurrence

Valentina Bayer Zubek  
Aureon Laboratories, Inc.  
Valentina.Bayer@aureon.com

David Verbel  
Aureon Laboratories, Inc.  
David.Verbel@aureon.com

Olivier Saidi  
Aureon Laboratories, Inc.  
Olivier.Saidi@aureon.com

## Abstract

*The task of predicting prostate specific androgen (PSA) recurrence following radical prostatectomy is important for the surveillance of patients with prostate cancer. This regression problem is complicated by the fact that data is censored, and there is no standard measurement of error for censored data. This paper applies modified regression trees, called Censored Time Trees (cTT™), to predict time to PSA recurrence. In order to assess the performance of cTT™, we explored different error measurements, such as the concordance index, AUC (area under the Receiver Operating Characteristic curve), sensitivity and specificity, and average error. cTT™ are compared against support vector regression, modified for censored data, and are found to have similar performance.*

## 1. Introduction

The estimated number of newly diagnosed cases of prostate cancer in the US is 230,000, and it is predicted that a male born today has a 16% chance of being diagnosed with prostate cancer. Prostate cancer has the second highest rate (4%) of cancer mortality in men, after lung cancer (50%). More than 75% of all prostate cancers occur in men over the age of 65 (nihseniorhealth.gov). For men that underwent radical prostatectomy, subsequent increase in PSA levels indicates that the cancer has spread outside the prostate and may lead to metastasis.

Our goal is to predict prostate specific androgen (PSA) recurrence post-prostatectomy. The challenge is to exploit information from **censored** patients, those who did not have recurrence by their last follow-up visit, as well as from **non-censored** patients, those for which we know the time when they experienced PSA recurrence.

In this paper, we introduce Censored Time Trees (cTT™) that extend regression trees [1] by taking into account the censored nature of survival data. Regression trees test features against some thresholds and have time predictions in the leaves. They are attractive because they can be summarized as simple rules and can be viewed as feature selectors. When applied to prostate cancer, the

time predictions provided by cTT™ can be seen as an advantage over the nomogram, which predicts the probability of a patient being PSA recurrence-free within 7 years [2].

Other authors have modified regression trees to apply them to survival data, creating the so-called survival trees. However, because Censored Time Trees™ predict time in the leaves, our approach is different from survival trees, which output hazard rates [3, 4] or Kaplan Meier estimates [5, 6]. In some papers that adapt regression trees to survival data, splitting of nodes is based on within-node homogeneity. For example, Davis and Anderson [3] use the negative log-likelihood of an exponential model, Gordon and Olshen [5] use distance measures between Kaplan Meier survival curves in the children of a node, and LeBlanc and Crowley [4] use the first step of a full likelihood estimation procedure for the Cox proportional hazards model. In other papers that depart from regression trees, between-node separation is chosen (e.g., Segal [6] employs two-sample log-rank test statistics).

In order to improve the time estimates, we employed bagging [7] to average the time predictions of several learned Censored Time Trees™. A recent paper by Hothorn et al. [8] employed bagging of survival trees to produce an aggregated Kaplan-Meier curve.

A challenge when working with censored data is defining a meaningful error for censored patients. This paper explores several error measurements and their shortcomings.

The paper first details the algorithms for constructing Censored Time Trees™ and support vectors for regression with censored data (SVRc™). Then it describes several error measurements, the patient data and experimental results, and ends with conclusions.

## 2. Censored time trees™

In regression, a data point consists of a vector of feature values and an observed time. For survival data, the observed time of a patient can be the actual event (e.g., PSA recurrence), or the date of the last follow-up. In the first case, the patient is non-censored, in the latter, he is censored.

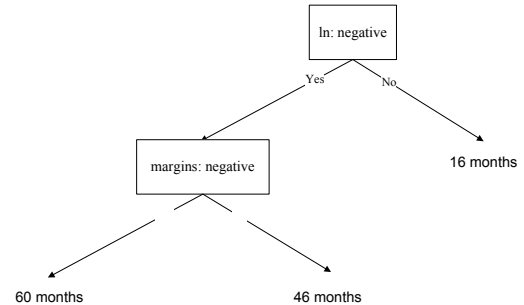
A regression tree [1] is a binary tree that tests one feature in each node, until a stopping criterion is met (e.g., the node has a small error or it has less than 20 patients). Depending on the result of the test, one of its branches will be followed. Each leaf of the tree makes a (constant value) time prediction for the patients that fall into that leaf by taking the average of their observed times. The best split in a node is the one that most reduces the cumulative risk of its children. In regression trees, all observed times are non-censored, and the risk is the mean square error between the observed and predicted times. Figure 1 shows a simple example of a regression tree that tests lymph nodes (ln) status and margins status and predicts time to PSA recurrence. The left branch is followed by patients for whom the test in the node is true, and the right branch by the patients for whom the test is false. For example, a patient that has positive lymph nodes gets a prediction of 16 months to PSA recurrence. A patient that has negative lymph nodes and positive margins gets a prediction of 46 months to PSA recurrence.

Censored Time Trees<sup>TM</sup> inherit the above characteristics from regression trees, but we modified the definition of risk in a node to be the exponential log-likelihood loss [3],  $R = D - D * \ln(D / Y)$ , where  $D$  is the number of non-censored patients in the node and  $Y$  is the sum of observed times for all patients in the node. We also experimented with a different formula for the risk, which modifies the mean square error such that censored patients whose predicted times are greater than their observed times contribute zero error (this is a slightly more comprehensive error definition than the one in [9]). The rationale for this modified error is that censored patients may recur at any time after their last follow-up. However, the results on internal cross-validation on the training dataset were better for the definition of risk as the exponential log-likelihood loss, and we will employ this definition in the paper.

To avoid overfitting, the learned tree is pruned using cost-complexity pruning ([1]). This method first grows a tree on the training data, from which it computes a sequence of pruning levels. Each pruning level will produce a pruned sub-tree of the original tree. The quality of a sub-tree is assessed through a cost-complexity measure that linearly combines the risk of the tree and its number of leaves. Cross-validation is employed to determine the pruning level  $\alpha_{\min}$  that achieves the minimum cost-complexity measure (see [1] for details). Finally, a chi-square test determines the smallest tree that has a cost-complexity measure close to that of the sub-tree corresponding to  $\alpha_{\min}$  ([3]).

## 2.1. Bagging censored time trees<sup>TM</sup>

A single tree has low bias (bias = systematic error) and high variance (variance = change in tree produced by change in training set). Bagging [7] can reduce variance. The method generates multiple versions of a predictor and then it averages their time estimates for a new patient. First,  $n$  bootstrap replicates ([10]) are drawn from the training data, with replacement, with sampling stratified by censored status. A Censored Time Tree<sup>TM</sup> is learned on each of the bootstrap replicates. The predicted time for a patient is the average of the predictions from the  $n$  Censored Time Trees<sup>TM</sup>.



**Figure 1. Simple example of regression tree/ Censored Time Tree<sup>TM</sup> that predicts time to PSA recurrence.**

## 3. Support vectors for regression with censored data (SVRc<sup>TM</sup>)

Support vector machines ([11, 12]) were originally developed for classification, their main advantage being their ability to handle high dimensional data. Support vector machines were extended for regression ([13]). Our company has patented SVRc<sup>TM</sup> ([14]), a support vector regression algorithm that is applied to survival data. The output of SVRc<sup>TM</sup> is also predicted time. For censored patients, SVRc<sup>TM</sup> assigns less of a penalty to over-predicted times than to under-predicted times, because the event did not happen before the last follow-up. For non-censored patients, SVRc<sup>TM</sup> assigns less of a penalty to under-predicted times than to over-predicted times, because the event may have happened before it was recorded.

## 4. Measurements

### 4.1. Concordance index

The Concordance Index (CI) [15] is the probability of predicting in the correct order the times for two patients, given that 1) both patients recurred, or 2) the patient that recurred has an earlier observed time than the last follow-

up time for the censored patient. Let  $t_1$  and  $t_2$  be the observed times for any two patients, such that  $t_1 < t_2$ , and let  $p_1$  and  $p_2$  be their predicted times. In case 1), both patients are non-censored. In case 2), the first patient is non-censored and the second patient is censored. In both cases, predicting the time in the correct order means that  $p_1 < p_2$ . The CI is an extension of the area under the ROC curve, AUC, which is discussed below, to survival data.

## 4.2. ROC, AUC, sensitivity and specificity

For binary classification, several measurements can be employed to assess the performance of classifiers, such as the ROC (receiver operating characteristic) curve, its area (AUC), sensitivity and specificity. The patients are labeled as belonging to the positive or negative class (usually, positive means sick, and negative means healthy). Binary classifiers predict that the patient belongs to the positive or negative class.

Regression tasks, in which both the observed times and the predicted times are real values, can be transformed into binary classes relative to a threshold. For example, for PSA recurrence the observed times are split relative to a threshold of 60 months, such that the **positive class contains all non-censored patients that recurred prior to 60 months** (early recurrence is worse), and the **negative class contains all patients with observed time greater than 60 months**. The rationale for this definition of classes is that for positive (sick) patients we wanted to be sure that they experienced PSA recurrence prior to the threshold, while patients that experienced PSA recurrence later than the threshold or those who were censored after the threshold (i.e., their last follow-up occurred later) may be considered negative (healthy) relative to that threshold (we at least know that they did not recur before the threshold). The predicted times are similarly split relative to a threshold from the range of predicted values (e.g., 55 months).

The **sensitivity** is defined as the true positive rate (i.e., the percentage of sick patients that are correctly predicted). The **specificity** is defined as the true negative rate (i.e., the percentage of healthy patients that are correctly predicted).

The **ROC curve** [16, 17] provides a visual interpretation of the tradeoff between sensitivity and specificity, by plotting the sensitivity vs. 1-specificity for different thresholds that split the predicted values into two classes. The **AUC** is defined as the probability that a (positive, negative) pair of patients is correctly ordered by the classifier [18, 19], which in our case means that the predicted time for the positive patient is less than for the negative patient. The AUC differs from the concordance index because it only compares pairs of patients from different classes (it does not compare pairs of positive

non-censored patients nor pairs of negative non-censored patients).

## 4.3. Average absolute error

We also report the absolute error between observed and predicted times, averaged over all patients. We report separately the average error for censored and non-censored patients.

For a non-censored patient, the observed time is the event time; therefore the absolute error between the observed time and the predicted time is well defined. However, for a censored patient, we only know that the event of interest (e.g., time to PSA recurrence) happens after the observed time of the last follow-up. Therefore the absolute difference between observed and predicted times underestimates the true error between event time and predicted time, when the predicted time is less than the observed time, and it can either underestimate or overestimate the true error when the predicted time is greater than the observed time.

## 5. Experiments

### 5.1. Censored data

The training data set consists of 262 patients that were treated at the Baylor College of Medicine; 37 of them were non-censored (14.1%). The test data set consists of 331 patients from Memorial Sloan-Kettering Cancer Center (MSKCC), 78 of which are non-censored (23.6%).

The two cohorts have 12 clinicopathological features in common; however two of the features (tumor node metastasis stage and the result of digital rectal examination) were measured differently by the two institutions and were therefore eliminated. Two additional features (age and biopsy dominant Gleason score) were eliminated because they did not correlate well with the observed times on the training data, based on the concordance index between the feature values and the observed times. There were 8 remaining features on which we ran the experiments. These features include results of biopsy (e.g., total Gleason grade) and prostatectomy (e.g., Gleason grade, status of lymph nodes, margins, seminal vesicle invasion and extra capsular extension) and a measurement of PSA level in blood before the surgery. The 8 selected features were measured in all patients (there were no missing values). Out of them, 4 are binary, one is real and 3 are ordinal.

### 5.2. Results

We compared the following algorithms: one cTT<sup>TM</sup>, bagging 25 cTT<sup>TM</sup>, and SVRc<sup>TM</sup>. The single Censored

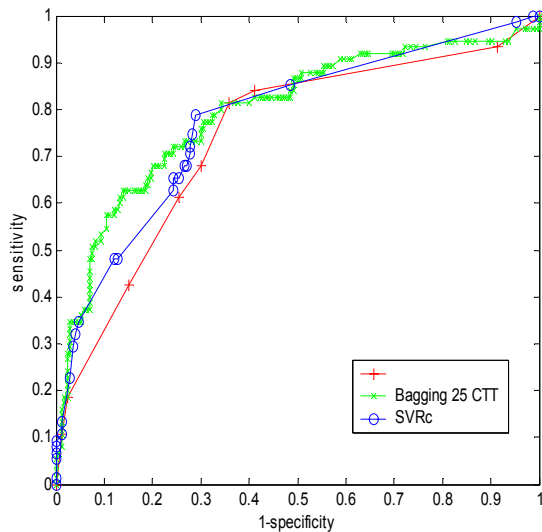
Time Tree<sup>TM</sup> is learned on the entire training data, as is the SVRc<sup>TM</sup>. For bagging, 25 bootstrap replicates with replacement, stratified by censored status, are drawn from the training data; on each bootstrap replicate a Censored Time Tree<sup>TM</sup> is learned, then for each patient the time estimates of the 25 cTT<sup>TM</sup> are averaged to obtain the time prediction. The prediction times are computed for patients in both training and test sets. No feature scaling is performed for cTT<sup>TM</sup>. For SVRc<sup>TM</sup>, the features are scaled between [-1, 1].

Table 1 presents the concordance indexes for the three algorithms. Bagging 25 cTT<sup>TM</sup> performed the best on the test dataset, better than one cTT<sup>TM</sup> (as expected), and only slightly better than SVRc<sup>TM</sup>.

**Table 1. Concordance index**

| Algorithm                    | Training set | Test set     |
|------------------------------|--------------|--------------|
| One cTT <sup>TM</sup>        | .8785        | .7167        |
| Bagging 25 cTT <sup>TM</sup> | .9386        | <b>.7679</b> |
| SVRc <sup>TM</sup>           | .8436        | .755         |

For the ROC and AUC analysis, the observed times (on both training and test data sets) were split into positive and negative classes relative to a threshold of 60 months, taking into account the censored status, as described in Section 4.2. The predicted times are split into classes relative to each distinct value from the predicted range, from which we compute the sensitivity and 1-specificity that define the operating points on the ROC curve.



**Figure 1. ROC curves on the test dataset**

Figure 1 displays the ROC curves on the test dataset for the three algorithms. The three ROC curves cross each

other, however most of the time the ROC curve for bagging is above the other two, and one explanation for this is that bagging cTT<sup>TM</sup> produces more unique predicted values, and therefore more operating points, than the other two algorithms (193 versus 26 for SVRc<sup>TM</sup> and 9 for one cTT<sup>TM</sup>).

Table 2 presents the AUC for the three algorithms. AUC is usually larger than the concordance index, and the rankings of the algorithms (bagging cTT<sup>TM</sup>, followed by SVRc<sup>TM</sup>, then one cTT<sup>TM</sup>) remains the same as for the concordance index.

**Table 2. Area under the ROC curve (AUC)**

| Algorithm                    | Training set | Test set     |
|------------------------------|--------------|--------------|
| One cTT <sup>TM</sup>        | .8876        | .7412        |
| Bagging 25 cTT <sup>TM</sup> | .9504        | <b>.7946</b> |
| SVRc <sup>TM</sup>           | .8425        | .7771        |

The experts advised us that for predicting PSA recurrence, sensitivity is more important than specificity, because we do not want to miss patients that need treatment. For this reason, we chose the threshold from the predicted times on the training data that maximizes specificity such that sensitivity is above 90%. Each algorithm has its own threshold. Table 3 presents the sensitivity and specificity for one cTT<sup>TM</sup>, bagging 25 cTT<sup>TM</sup> and SVRc<sup>TM</sup>. One cTT<sup>TM</sup> and SVRc<sup>TM</sup> have sensitivity close to 1 and small specificity on both training and test data sets, and this is probably explained by the fact that both algorithms have a small number of predicted values. The performance of bagging cTT<sup>TM</sup> was quite good on the test data, both sensitivity and specificity being around .7. However, if we chose other thresholds (e.g., such that the sensitivity and specificity on the training data are closest together), all three algorithms have more similar sensitivity and specificity on the test data.

**Table 3. Sensitivity and specificity on training and test datasets, computed at the threshold t (in months) that achieves the maximum specificity with sensitivity > 90% on training**

| Algorithm                    | Training set |      |      | Test set    |             |
|------------------------------|--------------|------|------|-------------|-------------|
|                              | t            | se   | sp   | se          | sp          |
| One cTT <sup>TM</sup>        | 62           | 1    | 0.11 | 0.93        | 0.09        |
| Bagging 25 cTT <sup>TM</sup> | 54           | 0.91 | 0.82 | <b>0.77</b> | <b>0.68</b> |
| SVRc <sup>TM</sup>           | 72           | 0.97 | 0.22 | 0.98        | 0.04        |

In terms of the features tested by the algorithms, SVRc<sup>TM</sup> selects only four features, three of which are also tested by more than 18 of the 25 bagged cTT<sup>TM</sup>.

Tables 4 and 5 display the average absolute error for all patients and its break-up for censored and non-censored patients, on training and test datasets. On the training data, both cTT<sup>TM</sup> methods have an average error of about 24 months, and SVRc<sup>TM</sup> errs by 27 months. On the test data, the average error for all algorithms drops to about 21 months. One would expect the error to increase from training to test datasets, and indeed when looking at non-censored patients, for whom we have actual recurrence times, we see that expected trend. However, on both datasets more than three quarters of the patients are censored, and their decrease in error from training to test data dictates the overall trend.

**Table 4. Average absolute errors (in months) on training data set**

| Algorithm                    | All patients | Censored    | Non-censored |
|------------------------------|--------------|-------------|--------------|
| One cTT <sup>TM</sup>        | 24.7         | 25.4        | 20.1         |
| Bagging 25 cTT <sup>TM</sup> | <b>24.2</b>  | <b>25.1</b> | <b>18.8</b>  |
| SVRc <sup>TM</sup>           | 26.9         | 26.7        | 28.7         |

**Table 5. Average absolute errors (in months) on test data set**

| Algorithm                    | All patients | Censored    | Non-censored |
|------------------------------|--------------|-------------|--------------|
| One cTT <sup>TM</sup>        | 21.4         | <b>19.4</b> | 27.7         |
| Bagging 25 cTT <sup>TM</sup> | <b>21.1</b>  | 19.5        | <b>26.5</b>  |
| SVRc <sup>TM</sup>           | 21.3         | 17.9        | 32.6         |

## 6. Conclusions

This paper introduces a modification of regression trees that output time predictions and applies them to survival data for predicting PSA recurrence. The performance of cTT<sup>TM</sup> improves with bagging, and it is slightly better than the performance of SVRc<sup>TM</sup>, for several error measurements including concordance index, AUC, and average error.

In future work we will apply Censored Time Trees<sup>TM</sup> to datasets with additional types of features, such as biomarkers, and will assess their contribution to the algorithm's performance.

## 7. Acknowledgments

We would like to thank Faisal Khan for tuning the parameters of SVRc<sup>TM</sup>, and Marina Sapir for her idea of eliminating features that do not correlate well with the observed times.

## 8. References

- [1] L. Breiman, L., J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees", Chapman & Hall/CRC, 1984.
- [2] M. Kattan, T. Wheeler, and P. Scardino, "Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer", *Journal of Clinical Oncology*, vol. 17, no. 5, 1999, pp. 1499-1507.
- [3] R. Davis, and J. Anderson, "Exponential survival trees", *Statistics in Medicine*, vol. 8, 1989, pp. 947-961.
- [4] M. LeBlanc, and J. Crowley, "Relative risk trees for censored survival data", *Biometrics* 48, 1992, pp. 411-425.
- [5] L. Gordon, and R. Olshen, "Tree-structured survival analysis", *Cancer treatment reports*, vol. 69, no. 10, 1985, pp. 1065-1068.
- [6] M. Segal, "Regression trees for censored data", *Biometrics* 44, 1988, pp.35-48.
- [7] L. Breiman, "Bagging predictors", *Machine Learning*, vol. 24, no. 2, 1996, pp. 123-140.
- [8] T. Hothorn, B. Lausen, A. Benner, and M. Radespiel-Troger, "Bagging survival trees", *Statistics in Medicine*, **23**(1), 2004, pp. 77-91.
- [9] O. Mangasarian, N. Street, and W. Wolberg, "Breast cancer diagnosis and prognosis via linear programming", Technical Report 94-10, University of Wisconsin, Madison, 1994.
- [10] B. Efron, and R. Tibshirani, "An introduction to the bootstrap", Chapman & Hall, New York, 1993.
- [11] V. Vapnik, "The nature of statistical theory", Springer Verlag, 1995.
- [12] N. Cristianini, and J. Shawe-Taylor, "An introduction to support vector machines", Cambridge University Press, 2000.
- [13] A. Smola, and B. Scholkopf, "A tutorial on support vector regression", NeuroCOLT2 Technical Report Series, NC2-TR-1998-030, 1998.
- [14] L. Yan, D. Verbel, and O. Saidi, "Predicting prostate cancer recurrence via maximizing the concordance index", ACM SIGKDD Conference 2004 Proceedings, pp. 479 - 485.
- [15] F. Harrell, R. Califf, and D. Pryor, "Evaluating the yield of medical tests", *Journal of American Medical Association*, vol. 247, 1982, pp. 2543-2546.

[16] F. Provost, and T. Fawcett, "Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions.", Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1997.

[17] S.H. Park, J. M. Goo, and C. Jo, "Receiver operating characteristic (ROC) curve: practical review for radiologists", Korean Journal of Radiology, 5(1), 2004, pp. 11-18.

[18] D. Hand, and R. Till, "A simple generalization of the area under the ROC curve for multiple class classification problems", Machine Learning, 45, 2001, pp. 171-186.

[19] J. Hanley, and B. McNeil, "The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve", Radiology, 143, 1982, pp. 29-36.