

# Machine Learning for Survival Analysis: A New Approach

---

**D. Dooling P. Green A. Kim D. Scroggin L. Stevens J. Webster**

*Innovative Oncology Business Solutions,  
4901 Lang Ave NE,  
Albuquerque, NM 87109, USA*

*E-mail:* [ddooling@innovativeobs.com](mailto:ddooling@innovativeobs.com), [pgreen@innovativeobs.com](mailto:pgreen@innovativeobs.com),  
[akim@innovativeobs.com](mailto:akim@innovativeobs.com), [dscroggin@innovativeobs.com](mailto:dscroggin@innovativeobs.com),  
[lstevens@innovativeobs.com](mailto:lstevens@innovativeobs.com), [jwebster@innovativeobs.com](mailto:jwebster@innovativeobs.com)

**ABSTRACT:** We have applied a little-known data transformation on subsets of the Surveillance, Epidemiology, and End Results (SEER) publically available data of the National Cancer Institute (NCI) to make it suitable input to standard machine learning classifiers. This transformation properly treats the right-censored data in the SEER data and the resulting Random Forest and Multi-Layer Perceptron models predict full survival curves. Treating the 6, 12, and 60 months points of the resulting survival curves as 3 binary classifiers, the 18 resulting classifiers have AUC values ranging from .765 to .885. Further evidence that the models have generalized well from the training data is provided by the extremely high levels of agreement between the random forest and neural network models predictions on the 6, 12, and 60 month binary classifiers.

---

## Contents

<b>1</b>	<b>Introduction and Background</b>	<b>1</b>
<b>2</b>	<b>Methodology</b>	<b>3</b>
2.1	Data acquisition	3
2.2	Data preparation and preprocessing	3
2.3	Colon Cancer Data	5
2.4	Lung Cancer Data	6
2.5	Breast Cancer Data	7
<b>3</b>	<b>Machine Learning Survival Analysis with Censored Data</b>	<b>9</b>
3.1	Survival Analysis	10
3.2	Transformation of Censored Data for Machine Learning	11
<b>4</b>	<b>Prediction Models</b>	<b>13</b>
4.1	Decision Trees and Random Forests	14
4.2	MLP Neural Networks	14
<b>5</b>	<b>Performance Metrics</b>	<b>14</b>
5.1	Model Agreement	14
<b>6</b>	<b>Web Applications</b>	<b>14</b>
<b>7</b>	<b>Further Directions</b>	<b>14</b>
<b>A</b>	<b>Selected Features</b>	<b>14</b>
A.1	Colon Cancer Feature Selection	15
A.2	Lung Cancer Feature Selection	17
A.3	Breast Cancer Feature Selection	20
A.4	Pseudocode for the Data Transformation	22
<b>B</b>	<b>Model Architecture and Python Code</b>	<b>22</b>
<b>C</b>	<b>GitHub Repositories</b>	<b>22</b>

---

## 1 Introduction and Background

Extracting actionable information from data is changing the fabric of modern business. A class of techniques that transforms data into actionable information goes by the name of Machine Learning [13]. Machine Learning has recently become a popular method to answer

questions and solve problems that are too complex to solve via traditional methods. The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) has been collecting data because intuitively researchers feel confident that this data is capturing information that has buried within it useful information in the form of relationships between the types of data collected (demographic as well as staging information) and the survival outcomes. Though this relationship evades capture by traditional methods, it is possible to surface it with the two machine learning techniques known as **Random Forests** and **Neural Networks**. These two methods produce very similar results when applied to the SEER dataset, and are based on two almost diametrically opposed learning philosophies, which lends confidence in the validity of the results.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is the most recognized authoritative source of information on cancer incidence and survival in the United States. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28 percent of the US population.

Quoting directly from the SEER website [11]:

The SEER program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. This program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data. The mortality data reported by SEER are provided by the National Center for Health Statistics. The population data used in calculating cancer rates is obtained periodically from the Census Bureau. Updated annually and provided as a public service in print and electronic formats, SEER data are used by thousands of researchers, clinicians, public health officials, legislators, policymakers, community groups, and the public.

One characteristic of the SEER data and that is shared by many datasets in the medical field goes by the name of "censored data." The SEER data contains the number of months each patient survived, as well as an indicator variable showing whether or not the patient is still alive at the end of the data collection period. Methods to deal effectively with this kind of "right-censored data" include Kaplan-Meier curves and Cox's Proportional Hazard models [4]. The Kaplan-Meier techniques only give estimates for cohorts of patients and are not applicable for predicting the survival curve for a single patient, and the Cox Proportional Hazard models require a fairly restrictive set of assumptions to be satisfied in order to yield reliable results. In addition, the Cox Proportional Hazard models are not able to capture the nonlinear relationships between the given data fields that go into making predictions; they can only capture the first-order linear relationships.

Previous work applying machine learning methods to subsets of the SEER data include creative attempts to deal with the problems presented by "right-censored data." The authors of [14] use semi-supervised learning techniques to predict 5 year survival, essentially

imputing values for SEER records where the survival months information is censored at a value less than 5 years. The authors of [15] investigate the effects of comorbidities; i.e., patients with two different cancer diagnoses, but their treatment of the censored data underestimates the survival probabilities. All records representing patients who survived at least 60 months as well as all those who died earlier than 60 months were considered, but patients alive prior to 60 months but censored out of the study before 60 months were not included. This treatment biases the data and the predictions, leading to overly pessimistic survival probabilities predicted by the trained models.

To overcome these limitations of the traditional methods, IOBS has applied a little-known technique to transform the SEER data to make it amenable to more powerful machine learning methods. The essential idea is to recast the problem to an appropriate discrete classification problem instead of a regression problem (predicting survival months). Treating months after diagnosis as just another discrete feature, the SEER data (or any other right-censored data) can be transformed simply so as to make predictions for the hazard function, probability of dying in the next month, given that the patient has not yet died. The full survival function can then be derived from the hazard function. Details of this transformation can be found in this blog post [1].

## 2 Methodology

### 2.1 Data acquisition

We used the publically available 1973-2012 SEER incidence data files corresponding to colon, breast and lung cancer contained in the following list. SEER requires that researchers submit a request for the data, which includes an agreement form. Detailed documentation explaining the contents of both the incidence data files used in this study as well as a data dictionary for the 1973-2012 SEER incidence data files are available without the need to register or submit a data request [12].

- incidence\yr1973\_2012.seer9\COLRECT.txt
- incidence\yr1973\_2012.seer9\BREAST.txt
- incidence\yr1973\_2012.seer9\RESPIR.txt
- incidence\yr1992\_2012.sj\_la\_rg\_ak\COLRECT.txt
- incidence\yr1992\_2012.sj\_la\_rg\_ak\BREAST.txt
- incidence\yr1992\_2012.sj\_la\_rg\_ak\RESPIR.txt
- incidence\yr2000\_2012.ca\_ky\_lo\_nj\_ga\COLRECT.txt
- incidence\yr2000\_2012.ca\_ky\_lo\_nj\_ga\BREAST.txt
- incidence\yr2000\_2012.ca\_ky\_lo\_nj\_ga\RESPIR.txt
- incidence\yr2005.lo\_2nd\_half\COLRECT.txt
- incidence\yr2005.lo\_2nd\_half\BREAST.txt
- incidence\yr2005.lo\_2nd\_half\RESPIR.txt

### 2.2 Data preparation and preprocessing

A great deal of data munging is necessary before using these SEER incidence files as input into machine learning algorithms. A preprocessing step common to each of three cancer

Code	Description
1	Male
2	Female

**Table 1.** Encoding of gender in the SEER incidence files. These types of categorical variables need to be transformed via one-hot-encoding.

types studied involves the `STATE-COUNTY RECODE`. The `STATE-COUNTY RECODE` field is a state-county combination where the first two characters represent the state FIPS code and the last three digits represent the FIPS county code. This particular field illustrates an important feature of machine learning, that between *categorical features* and *numeric features*. All input into a machine learning algorithm must be numeric, but real numbers carry with them the usually extremely useful property known as the well-ordering property of the real numbers. But if one is tasked with encoding a categorical feature into suitable numeric format for machine learning, it is necessary to do so in a way that removes the well-ordering property [2].

As a simple example of how to correctly treat categorical variables in a machine learning context, consider the SEER variable `SEX`. This variable is encoded with a numeric 1 for males and a numeric 2 for females as shown in Table 1. Values such as "Male" and "Female" encoded as numbers are dangerous because if not handled properly, they can generate bogus results [7]. The proper way to transform the SEER `SEX` variable is to create two additional variables: `sex_Male` and `sex_Female`, and then to eliminate the variable `SEX`. For example,

$$\begin{array}{c|c} \text{Sex} & \\ \hline 1 & \end{array} \longrightarrow \begin{array}{c|c|c} \text{sex\_Male} & \text{sex\_Female} & \\ \hline 1 & 0 & \end{array} \quad (2.1)$$

and

$$\begin{array}{c|c} \text{Sex} & \\ \hline 2 & \end{array} \longrightarrow \begin{array}{c|c|c} \text{sex\_Male} & \text{sex\_Female} & \\ \hline 0 & 1 & \end{array} \quad (2.2)$$

The procedure outlined in Equations (2.1, 2.2) needs to be applied to all of the nominal categorical variables in the SEER data that we wish to include in our predictive models. In particular, in order to include the geographical information contained in the SEER categorical variable `STATE-COUNTY RECODE`, it becomes necessary to create a new feature variable for each of the distinct (state,county) pairs in the data. In the United States, there are approximately 3,000 counties. Clearly, transforming the `STATE-COUNTY RECODE` data representation into distinct (state\_county) columns will explode the data to become wider than is optimal for machine learning. Adding extra columns to your dataset, making it wider, requires more data rows (making it taller) in order for machine learning algorithms to effectively learn [2]. Because one-hot coding `STATE-COUNTY RECODE` would cause such

STATE-COUNTY RECODE	address	elevation	lat	lng
35001	Bernalillo+county+NM	5207.579772	35.017785	-106.629130
35003	Catron+county+NM	8089.242628	34.151517	-108.427605
35005	Chaves+county+NM	3559.931671	33.475739	-104.472330
35006	Cibola+county+NM	6443.415570	35.094756	-107.858387
35007	Colfax+county+NM	6147.749089	36.579976	-104.472330

**Table 2.** Example of the transformation of `STATE-COUNTY RECODE` to `elevation`, `lat`, and `lng`.

drastic shape changes in our data, we wish to avoid doing so. Fortunately, this variable, though given as a categorical variable, is actually an ordinal variable. There is an ordering among the (`state_county`) columns, name longitude, latitude, and elevation. We can transform the data in `STATE-COUNTY RECODE` into three new numerical columns: `lat`, `lng`, and `elevation`.

For example, Table (2) shows how five entries of `STATE-COUNTY RECODE` corresponding to counties within New Mexico would can be represented by the `elevation`, `lat`, and `lng` features.

It is a simple exercise to construct the full lookup table from the SEER `STATE-COUNTY RECODE` variable to the corresponding three values `elevation`, `lat`, and `lng`. Using the publically available `dafafile` from the United States Census Bureau [3] to construct query strings like the values of the `address` field in Table (2), it is possible to then programmatically query the Google Maps Geocoding API for the latitude and longitude [6], and the Google Maps Elevation API for the corresponding elevation [5]. An added benefit of this shift from the single categorical variable `STATE-COUNTY RECODE` to the three continuous numerical variables `lat`, `lng`, and `elevation` is that input into the web applications described later are not restricted to the states and counties covered in the SEER registries. The full lookup table analogous to Table 2 is available from a GitHub repository containing supplemental information for this study [8].

### 2.3 Colon Cancer Data

In this section we describe the data processing steps that were specific to the colon cancer model development. The four `COLRECT.txt` files were imported into a pandas `DataFrame` object. This data was then filtered according to the conditions in Table 3.

The following categorical features were one-hot encoded as described in section 2.2:

- `SEX`,
- `MARITAL STATUS AT DX`,
- `RACE/ETHNICITY`,
- `SPANISH/HISPANIC ORIGIN`,
- `GRADE`,

Column	Filter
SEQUENCE NUMBER-CENTRAL	$\neq$ "Unspecified"
AGE AT DIAGNOSIS	$\neq$ "Unknown age"
BIRTHDATE-YEAR	$\neq$ "Unknown year of birth"
YEAR OF DIAGNOSIS	$\geq 2004$
SURVIVAL MONTHS FLAG	$=$ "1"
CS TUMOR SIZE EXT/EVAL	$\neq$ ""
CS TUMOR SIZE	$\neq 999$
SEER RECORD NUMBER	$= 1$
PRIMARY SITE	$=$ "LARGE INTESTINE, (EXCL. APPENDIX)"
SEQUENCE NUMBER-CENTRAL	$= 0$

**Table 3.** Filters applied to the Colon Cancer data.

- PRIMARY SITE ,
- LATERALITY ,
- SEER HISTORIC STAGE A ,
- HISTOLOGY RECODE-BROAD GROUPINGS ,
- MONTH OF DIAGNOSIS ,
- VITAL STATUS RECODE .

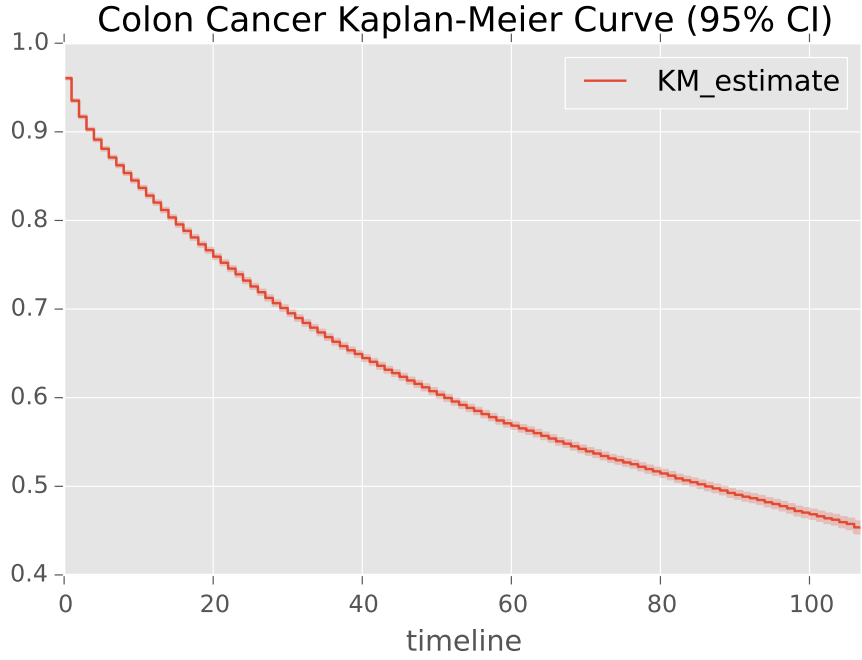
The STATE-COUNTY RECODE variable was dropped and replaced with the `elevation` , `lat` , and `lng` variables as illustrated in Table 2.

With just the above data preparation, it is possible to construct traditional Kaplan-Meier estimates of the survival curves for the colon cancer population represented by this subset of the data. After the above one-hot encoding procedure, the new variable `vital_status_recode_Dead` indicates that the patient is deceased if this variable  $= 1$ , or else that the patient's record is right-censored if this variable  $= 0$ . `SURVIVAL MONTHS` and `vital_status_recode_Dead` are all that is needed to construct the Kaplan-Meier estimate shown in Figure (1).

## 2.4 Lung Cancer Data

In this section we describe the data processing steps that were specific to the lung cancer model development. The four RESPIR.txt files were imported into a pandas DataFrame object. This data was then filtered according to the conditions in Table 4. The same list of categorical features as in the colon cancer case were then one-hot encoded.

With just the above data preparation, it is possible to construct traditional Kaplan-Meier estimates of the survival curves for the colon cancer population represented by this subset of the data. After the above one-hot encoding procedure, the new variable `vital_status_recode_Dead` indicates that the patient is deceased if this variable  $= 1$ , or else that the patient's record is right-censored if this variable  $= 0$ . `SURVIVAL MONTHS` and



**Figure 1.** Traditional Kaplan-Meier estimate of the survival curve for all colon cancer patients. Fitted with 113072 observations, 71804 censored.

Column	Filter
SEQUENCE NUMBER-CENTRAL	$\neq$ "Unspecified"
AGE AT DIAGNOSIS	$\neq$ "Unknown age"
BIRTHDATE-YEAR	$\neq$ "Unknown year of birth"
YEAR OF DIAGNOSIS	$\geq$ 2004
SURVIVAL MONTHS FLAG	= "1"
CS TUMOR SIZE EXT/EVAL	$\neq$ ""
CS TUMOR SIZE	$\neq$ 999
SEER RECORD NUMBER	= 1
PRIMARY SITE	= "LUNG & BRONCHUS"
SEQUENCE NUMBER-CENTRAL	= 0

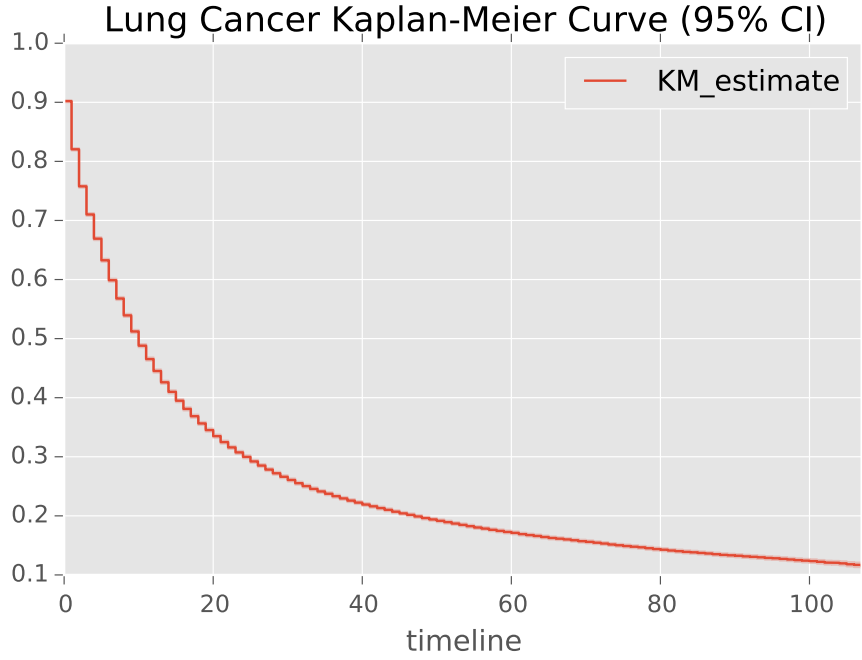
**Table 4.** Filters applied to the Lung Cancer data.

`vital_status_recode_Dead` are all that is needed to construct the Kaplan-Meier estimate shown in Figure (2).

## 2.5 Breast Cancer Data

In this section we describe the data processing steps that were specific to the lung cancer model development. The four BREAST.txt files were imported into a pandas DataFrame





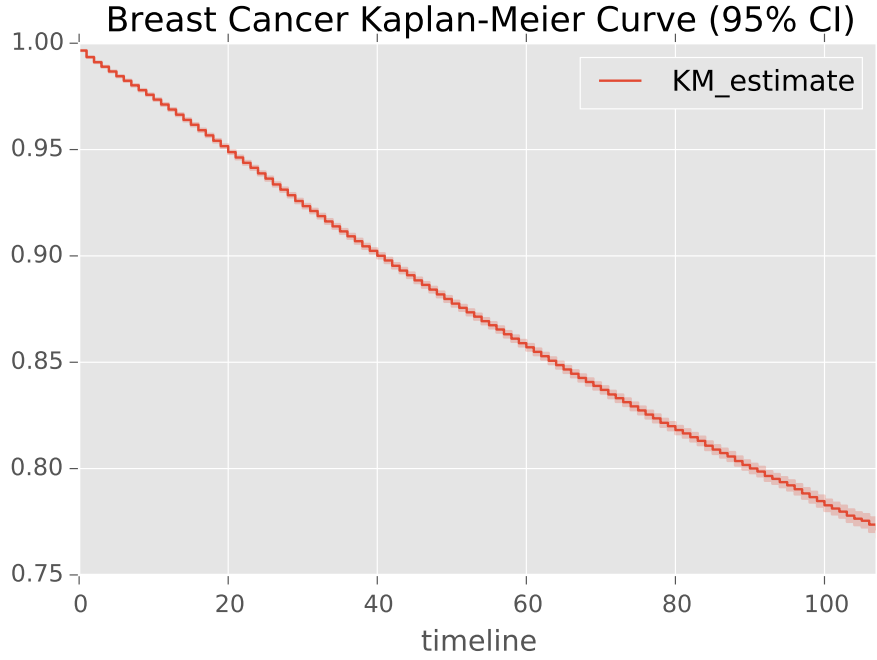
**Figure 2.** Traditional Kaplan-Meier estimate of the survival curve for all lung cancer patients. Fitted with 177089 observatins, 47409 censored.

Column	Filter
SEQUENCE NUMBER-CENTRAL	$\neq$ "Unspecified"
AGE AT DIAGNOSIS	$\neq$ "Unknown age"
BIRTHDATE-YEAR	$\neq$ "Unknown year of birth"
YEAR OF DIAGNOSIS	$\geq$ 2004
SURVIVAL MONTHS FLAG	$=$ "1"
CS TUMOR SIZE EXT/EVAL	$\neq$ " "
CS TUMOR SIZE	$\neq$ 999
SEER RECORD NUMBER	$=$ 1
SEQUENCE NUMBER-CENTRAL	$=$ 0

**Table 5.** Filters applied to the Breast Cancer data.

object. This data was then filtered according to the conditions in Table 5. The same list of categorical features as in the colon cancer case were then one-hot encoded.

With just the above data preparation, it is possible to construct traditional Kaplan-Meier estimates of the survival curves for the colon cancer population represented by this subset of the data. After the above one-hot encoding procedure, the new variable `vital_status_recode_Dead` indicates that the patient is deceased if this variable  $=$  1, or



**Figure 3.** Traditional Kaplan-Meier estimate of the survival curve for all breast cancer patients. Fitted with 329949 observatins, 292279 censored.

else that the patient’s record is right-censored if this variable = 0. `SURVIVAL MONTHS` and `vital_status_recode_Dead` are all that is needed to construct the Kaplan-Meier estimate shown in Figure (3).

Before applying machine learning models trained with these data sets, we review in section (3) the sailent features of survival analysis and censored data. We then describe in detail a method that takes full advantage of all the data, including the right-censored data, and which involves a simple and intuitive transformation, culminating in the full set of features and target variables listed in sections (A.1, A.2, A.3).

### 3 Machine Learning Survival Analysis with Censored Data

The above Kaplan-Meier estimates of the survival curves for colon (Figure (1), lung (Figure (2), and breast cancer (Figure (3) are constructed from the full population of cancer patients in the respective datasets. An unsatisfactory consequence is that these estimates are highly course-grained, and not very meaningful to an individual. Patients with very disparate characteristics are given the same prognosis by these Kaplan-Meier survival curve estimates. Therefore it is desirable to find robust predictors for survival curves of individual where the input is an individual record where the predictors are trained on larger populations.

### 3.1 Survival Analysis

To understand survival analysis, you first have to understand survival data - that survival times are *intervals* between certain kinds of events, that these intervals are often affected by a peculiar kind of "partial missingness" called *censoring*, and that censored data must be analyzed in a special way to avoid biased estimates and incorrect conclusions.

In the case of the SEER data, the starting point of the time interval is the diagnosis date. Even though survival times are continuous or nearly continuous numerical quantities, they're never almost never normally distributed. If non-normality were the only problem with survival data, you would be able to summarize survival times as medians and centiles instead of means and standard deviations, and you could compare survival between groups with nonparametric Mann-Whitney and Kruksal-Wallis testse instead of t tests and ANOVAs. But time-to-event data is susceptible to a special situation called *censoring*, which the usual parametric and non-parametric methods cannot handle. Therefore special methods have been developed to analyze censored data properly.

With survival data, including the SEER data considered in this study, you may not know the exact time of death for some subjects. Some of the SEER subjects are still alive at the the time of the latest SEER data release. When the `VITAL STATUS RECODE` variable indicates that the subject is still alive, the `SURVIVAL MONTHS` variable is only a lower bound on the true number of survival months; this is called the *date of last contact* mode of censoring.

Survival analysis allows us to model the time until an event happens. Time-to-event measures pose unique problems for the analyst. Suppose that you want to predict the survival time for patients receiving an experimental cancer treatment. After three years, some of the patients in the study have died, and you can compute the survival time for each of these patients. However, many of the patients are still living at the end of three years; you do not know their ultimate survival time. Statisticians call this problem *censoring*, a problem that surfaces when you try to model time-to-event reopnse measures using data captured over a limited time period.

The two kinds of censoring are right censoring and left censoring. If you only know that the pertinent event is *after* some date, as is the case for patients in the preceding example who survive to the end of the study, the data is right-censored. On the other hand, if you only know that the beginning of the pertinent time-to-event took place before a certain date, the data is left-censored. For example, if you know that every patient in the study received the experimental treatment before the study started but do not know the exact date of treatment, the data is left-censored. Data can be both right-censored and left-censored.

*Survival analysis* is a family of techniques developed to work with censored time-to-event response measures. Note that if censoring is not present, you may be able to model time-to-event using standard modeling techniques. For some studies, however, you would have to wait a very long time before every sampled observation has a terminal event; in the case of the experimental cancer treatment, some patients might live another 20 years. Hence, survival analysis techniques enable the analyst to take full advantage of available

data without waiting until every treated patient dies, every sampled part fails, or every tracked account closes.

The fundamental concept in survival analysis is the **survival curve**,  $S(t)$ , which is a

$$S(t) = 1 - CDF(t) \quad (3.1)$$

where  $CDF(t)$  is the probability of a lifetime less than or equal to  $t$ . From the survival curve we can derive the **hazard function**; for pregnancy lengths, hazard function maps from a time,  $t$ , to the fraction of pregnancies that continue until  $t$

$$\lambda(t) = \frac{S(t) - S(t+1)}{S(t)} \quad (3.2)$$

The numerator is the fraction of lifetimes that end at  $t$ , which is also  $PMF(t)$ . If someone gives you the CDF of lifetimes, it is easy to compute the survival and hazard functions. But in many real-world scenarios, we can't measure the distribution of lifetimes directly. We have to infer it. For example, suppose you are following a group of patients to see how long they survive after diagnosis. Not all patients are diagnosed on the same day, so at any point in time, some patients have survived longer than others. If some patients have died, we know their survival

If we wait until all patients are dead, we can compute the survival curve, but if we are evaluating the effectiveness of a new treatment, we can't wait that long! We need a way to estimate survival curves using incomplete information.

The general idea is that we can use the data to estimate the hazard function, then convert the hazard function to a survival curve. Once we have the hazard function, we can estimate the survival curve. The chance of surviving past time  $t$  is the chance of surviving all times up through  $t$ , which is the cumulative product of the complementary hazard function:

$$[1 - \lambda(0)][1 - \lambda(1)] \cdots [1 - \lambda(t)] \quad (3.3)$$

### 3.2 Transformation of Censored Data for Machine Learning

In this section we describe an intuitive way to transform right-censored data appropriately so that it may be used as input to machine learning algorithms that learn the hazard function described in section 3.1. The full details of this transformation, and a large inspiration for this study, can be found in this blog post [1].

The overall philosophy of the Kaplan-Meier estimate of the survival curve for a population differs fundamentally from the methods described below and used in this study. The Kaplan-Meier estimate of the survival curve is given by

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i} \quad (3.4)$$

where  $d_i$  are the number of death events at time  $t$  and  $n_t$  is the number of subjects at risk of death just prior to time  $t$ . Equation 3.4 uses the entire data set to arrive at an estimate of the entire population survival curve. In contrast, the method described below uses the

	cs_tumor_size	year_of_birth	survival_months	vital_status_recode_Death
newindex				
205	60	1951	3	1

**Table 6.** Example of four columns in an uncensored record in the untransformed dataset.

	cs_tumor_size	year_of_birth	survival_months	vital_status_recode_Death
newindex				
205	40	1950	3	0

**Table 7.** Example of four columns in a censored record in the untransformed dataset.

entire data set to learn a model so as to predict hazard and survival curves for all of the individual records in the data set.

The key observation is to note that the hazard function in Equation 3.2 can be readily learned via machine learning methods. It can be rewritten as

$$h(\mathbf{X}, t) = P(Y = t | Y \geq t, \mathbf{X}), \quad (3.5)$$

the probability that, if someone has survived up until month  $t$ , they will die in that month. where  $\mathbf{X}$  represents all of the data for that particular record, and in our case  $Y$  represents the true, uncensored number of survival months of the patient. What is actually provided in the SEER data is the related variable `SURVIVAL MONTHS`  $T$  (how long each subject was in the study), and whether they exited by dying or being censored ( $D$ ), `VITAL STATUS RECODE`.  $D$  is a Boolean variable, so  $D = 1$  if  $T = Y$ , and  $D = 0$  if  $T < Y$ .

Treating  $T$  is just another covariate is the key to the transformation. Each datapoint in the hidden classification problem is the combination of an  $\mathbf{X}_i$  in the original dataset plus some month  $t$ , and the classification problem is "did point  $\mathbf{X}_i$  die in month  $t$ ." We will call this new variable  $D_{it}$  (`newtarget`). We can transform our original data set into a new one, with one row for each month that each  $\mathbf{X}_i$  is in the sample; train a standard classifier on this new dataset with  $D_{it}$  as the target, and derive a survival model from the original dataset. Pseudocode for this transformation is found in section A.4.

Explicit examples will help make this transformation clear. The untransformed datapoint represented Table (6) is transformed to the multiple records shown in Table (8). All uncensored data is transformed in this way. All censored data is similarly transformed. The untransformed datapoint represented Table (7) is transformed to the multiple records shown in Table (9).

One obvious side effect of this transformation is that it explodes the data size. For this study, the original, untransformed colon cancer DataFrame has shape (113072, 106), and the total transformed colon cancer DataFrame has shape (4165251, 106). Similarly, the original, untransformed lung cancer DataFrame has shape (177089, 118), and the total

	<code>cs_tumor_size</code>	<code>year_of_birth</code>	<code>month</code>	<code>newtarget</code>
<code>newindex</code>				
205	60	1951	0	0
205	60	1951	1	0
205	60	1951	2	0
205	60	1951	3	1

**Table 8.** Example of four columns in an uncensored record in the transformed dataset.

	<code>cs_tumor_size</code>	<code>year_of_birth</code>	<code>month</code>	<code>newtarget</code>
<code>newindex</code>				
205	40	1950	0	0
205	40	1950	1	0
205	40	1950	2	0
205	40	1950	3	0

**Table 9.** Example of four columns in a censored record in the transformed dataset.

transformed colon cancer DataFrame has shape (3079931,118). The biggest explosion in data size occurred with the breast cancer data. The original, untransformed breast cancer DataFrame has shape (329949,70), and the total transformed breast cancer DataFrame has shape (15085711,70). Training machine learning algorithms on such large datasets, even after splitting into training and testing sets described below, require large RAM. All computations for this study were performed on a Dell XPS 8700 Desktop with 32GB of RAM.

## 4 Prediction Models

With the datasets transformed as described in section (3.2), we are now able to split them into training and testing sets in the usual manner. The classifier models described in this section are learning the hazard function: given all of the data given in sections (A.2, A.1, A.3), which includes the field `months` (the months after diagnosis), the models predict the target variable `newtarget`, which represents the probability of dying in that month, given that the patient represented by the record has survived up to that month. This prediction task should not be confused with the regression problem of trying to predict precisely in what month a patient will die. The hazard functions thus learned and predicted are intermediary products; what we are really pursuing are the survival functions for each patient that are derived from the learned and predicted hazard functions. From the resulting hazard functions for each unique patient, we can construct the resulting survival functions as presented in section (A.4) and explicitly given in python code in the notebooks at the github repository containing supplemental material for this study [8].

Model	6 Months AUC	12 Months AUC	60 Months AUC
Breast RF	.846	.885	.844
Breast NN	.855	.867	.836
Colon RF	.804	.806	.828
Colon NN	.797	.804	.841
Lung RF	.772	.796	.874
Lung NN	.765	.796	.875

**Table 10.** AUC values for the Random Forest and Neural Networks model binary classifiers derived from the full survival curve predictions; see text for details.

Cancer Type	% agreement 6 months	% agreement 12 months	% agreement 60 months
Colon	.981	.971	.915
Breast	.994	.984	.938
Lung	.861	.883	.900

**Table 11.** Percentage agreement for the Random Forest and Neural Network classifiers for 6, 12, and 60 month survival predictions on the test data for each cancer type.

#### 4.1 Decision Trees and Random Forests

#### 4.2 MLP Neural Networks

### 5 Performance Metrics

#### 5.1 Model Agreement

TO DO: Check if comorbidities are contributing to the outliers in the agreement boxplots that follow. Could mesh with the previous work [15].

### 6 Web Applications

CF this guy <http://kmplot.com/analysis/index.php?p=service&cancer=lung>

### 7 Further Directions

Discussion of causality. A certain Marital status is not a "cause" of a better prognosis; c.f. Simpson's Paradox. Implementation of Judea Pearl's Causality Calculus.

### A Selected Features

In this Appendix we explicitly list the features chosen for each of the Colon, Breast and Lung cancer predictive models. For each cancer type, the features chosen for the random forest and neural network models were the same, so as to be best able to compare the two

models. IPython notebooks explicitly providing all code, as well as html versions of the notebooks, are available from a GitHub repository providing supplemental material for this study [8].

### A.1 Colon Cancer Feature Selection

The feature set used as input into both the Random Forest and Neural Network models, after the transformation described in Section is given below and also available in full detail in the file `NewPatientColonML.html`.

- `cs_tumor_size`
- `elevation`
- `grade_cell type not determined`
- `grade_moderately differentiated`
- `grade_poorly differentiated`
- `grade_undifferentiated; anaplastic`
- `grade_well differentiated`
- `histology_recode_broad_groupings_acinar cell neoplasms`
- `histology_recode_broad_groupings_adenomas and adenocarcinomas`
- `histology_recode_broad_groupings_blood vessel tumors`
- `histology_recode_broad_groupings_complex epithelial neoplasms`
- `histology_recode_broad_groupings_complex mixed and stromal neoplasms`
- `histology_recode_broad_groupings_cystic, mucinous and serous neoplasms`
- `histology_recode_broad_groupings_ductal and lobular neoplasms`
- `histology_recode_broad_groupings_epithelial neoplasms, NOS`
- `histology_recode_broad_groupings_fibromatous neoplasms`
- `histology_recode_broad_groupings_germ cell neoplasms`
- `histology_recode_broad_groupings_lipomatous neoplasms`
- `histology_recode_broad_groupings_miscellaneous bone tumors`
- `histology_recode_broad_groupings_myomatous neoplasms`
- `histology_recode_broad_groupings_neuroepitheliomatous neoplasms`
- `histology_recode_broad_groupings_nevi and melanomas`
- `histology_recode_broad_groupings_paragangliomas and glomus tumors`
- `histology_recode_broad_groupings_soft tissue tumors and sarcomas, NOS`
- `histology_recode_broad_groupings_squamous cell neoplasms`
- `histology_recode_broad_groupings_synovial-like neoplasms`
- `histology_recode_broad_groupings_transitional cell papillomas and carcinomas`
- `histology_recode_broad_groupings_unspecified neoplasms`
- `lat`
- `laterality_Left: origin of primary`
- `laterality_Not a paired site`
- `laterality_Only one side involved, right or left origin unspecified`
- `laterality_Paired site, but no information concerning laterality; midline tumor`
- `laterality_Right: origin of primary`



- lng
- marital\_status\_at\_dx\_Divorced
- marital\_status\_at\_dx\_Married (including common law)
- marital\_status\_at\_dx\_Separated
- marital\_status\_at\_dx\_Single (never married)
- marital\_status\_at\_dx\_Unknown
- marital\_status\_at\_dx\_Unmarried or domestic partner
- marital\_status\_at\_dx\_Widowed
- month\_of\_diagnosis\_Apr
- month\_of\_diagnosis\_Aug
- month\_of\_diagnosis\_Dec
- month\_of\_diagnosis\_Feb
- month\_of\_diagnosis\_Jan
- month\_of\_diagnosis\_Jul
- month\_of\_diagnosis\_Jun
- month\_of\_diagnosis\_Mar
- month\_of\_diagnosis\_May
- month\_of\_diagnosis\_Nov
- month\_of\_diagnosis\_Oct
- month\_of\_diagnosis\_Sep
- number\_of primaries
- race\_ethnicity\_Amerian Indian, Aleutian, Alaskan Native or Eskimo
- race\_ethnicity\_Asian Indian
- race\_ethnicity\_Asian Indian or Pakistani
- race\_ethnicity\_Black
- race\_ethnicity\_Chinese
- race\_ethnicity\_Fiji Islander
- race\_ethnicity\_Filipino
- race\_ethnicity\_Guamanian
- race\_ethnicity\_Hawaiian
- race\_ethnicity\_Hmong
- race\_ethnicity\_Japanese
- race\_ethnicity\_Kampuchean
- race\_ethnicity\_Korean
- race\_ethnicity\_Laotian
- race\_ethnicity\_Melanesian
- race\_ethnicity\_Micronesian
- race\_ethnicity\_New Guinean
- race\_ethnicity\_Other
- race\_ethnicity\_Other Asian
- race\_ethnicity\_Pacific Islander
- race\_ethnicity\_Pakistani
- race\_ethnicity\_Polynesian

- race\_ethnicity\_Samoan
- race\_ethnicity\_Thai
- race\_ethnicity\_Tongan
- race\_ethnicity\_Unknown
- race\_ethnicity\_Vietnamese
- race\_ethnicity\_White
- seer\_historic\_stage\_a\_Distant
- seer\_historic\_stage\_a\_In situ
- seer\_historic\_stage\_a\_Localized
- seer\_historic\_stage\_a\_Regional
- seer\_historic\_stage\_a\_Unstaged
- sex\_Female
- spanish\_hispanic\_origin\_Cuban
- spanish\_hispanic\_origin\_Dominican Republic
- spanish\_hispanic\_origin\_Mexican
- spanish\_hispanic\_origin\_Non-Spanish/Non-hispanic
- spanish\_hispanic\_origin\_Other specified Spanish/Hispanic origin (excludes Dominican Republic)
- spanish\_hispanic\_origin\_Puerto Rican
- spanish\_hispanic\_origin\_South or Central American (except Brazil)
- spanish\_hispanic\_origin\_Spanish surname only
- spanish\_hispanic\_origin\_Spanish, NOS; Hispanic, NOS; Latino, NOS
- spanish\_hispanic\_origin\_Unknown whether Spanish/Hispanic or not
- year\_of\_birth
- year\_of\_diagnosis
- month
- newtarget

## A.2 Lung Cancer Feature Selection

The feature set used as input into both the Random Forest and Neural Network models, after the transformation described in Section is given below and also available in full detail in the file `NewPatientLungML.html` .

- cs\_tumor\_size
- elevation
- grade\_cell type not determined
- grade\_moderately differentiated
- grade\_poorly differentiated
- grade\_undifferentiated; anaplastic
- grade\_well differentiated
- histology\_recode\_broad\_groupings\_acinar cell neoplasms
- histology\_recode\_broad\_groupings\_adenomas and adenocarcinomas
- histology\_recode\_broad\_groupings\_blood vessel tumors

- histology\_recode\_broad\_groupings\_complex epithelial neoplasms
- histology\_recode\_broad\_groupings\_complex mixed and stromal neoplasms
- histology\_recode\_broad\_groupings\_cystic, mucinous and serous neoplasms
- histology\_recode\_broad\_groupings\_ductal and lobular neoplasms
- histology\_recode\_broad\_groupings\_epithelial neoplasms, NOS
- histology\_recode\_broad\_groupings\_fibroepithelial neoplasms
- histology\_recode\_broad\_groupings\_fibromatous neoplasms
- histology\_recode\_broad\_groupings\_germ cell neoplasms
- histology\_recode\_broad\_groupings\_gliomas
- histology\_recode\_broad\_groupings\_granular cell tumors & alveolar soft part sarcomas
- histology\_recode\_broad\_groupings\_lipomatous neoplasms
- histology\_recode\_broad\_groupings\_miscellaneous bone tumors
- histology\_recode\_broad\_groupings\_miscellaneous tumors
- histology\_recode\_broad\_groupings\_mucoepidermoid neoplasms
- histology\_recode\_broad\_groupings\_myomatous neoplasms
- histology\_recode\_broad\_groupings\_myxomatous neoplasms
- histology\_recode\_broad\_groupings\_nerve sheath tumors
- histology\_recode\_broad\_groupings\_neuroepitheliomatous neoplasms
- histology\_recode\_broad\_groupings\_nevi and melanomas
- histology\_recode\_broad\_groupings\_osseous and chondromatous neoplasms
- histology\_recode\_broad\_groupings\_parangliomas and glomus tumors
- histology\_recode\_broad\_groupings\_soft tissue tumors and sarcomas, NOS
- histology\_recode\_broad\_groupings\_squamous cell neoplasms
- histology\_recode\_broad\_groupings\_synovial-like neoplasms
- histology\_recode\_broad\_groupings\_thymic epithelial neoplasms
- histology\_recode\_broad\_groupings\_transitional cell papillomas and carcinomas
- histology\_recode\_broad\_groupings\_trophoblastic neoplasms
- histology\_recode\_broad\_groupings\_unspecified neoplasms
- lat
- laterality\_Bilateral involvement, lateral origin unknown; stated to be single primary
- laterality\_Left: origin of primary
- laterality\_Not a paired site
- laterality\_Only one side involved, right or left origin unspecified
- laterality\_Paired site, but no information concerning laterality; midline tumor
- laterality\_Right: origin of primary
- lng
- marital\_status\_at\_dx\_Divorced
- marital\_status\_at\_dx\_Married (including common law)
- marital\_status\_at\_dx\_Separated
- marital\_status\_at\_dx\_Single (never married)
- marital\_status\_at\_dx\_Unknown
- marital\_status\_at\_dx\_Unmarried or domestic partner

- marital\_status\_at\_dx\_Widowed
- month\_of\_diagnosis\_Apr
- month\_of\_diagnosis\_Aug
- month\_of\_diagnosis\_Dec
- month\_of\_diagnosis\_Feb
- month\_of\_diagnosis\_Jan
- month\_of\_diagnosis\_Jul
- month\_of\_diagnosis\_Jun
- month\_of\_diagnosis\_Mar
- month\_of\_diagnosis\_May
- month\_of\_diagnosis\_Nov
- month\_of\_diagnosis\_Oct
- month\_of\_diagnosis\_Sep
- number\_of primaries
- race\_ethnicity\_Amerian Indian, Aleutian, Alaskan Native or Eskimo
- race\_ethnicity\_Asian Indian
- race\_ethnicity\_Asian Indian or Pakistani
- race\_ethnicity\_Black
- race\_ethnicity\_Chamorroan
- race\_ethnicity\_Chinese
- race\_ethnicity\_Fiji Islander
- race\_ethnicity\_Filipino
- race\_ethnicity\_Guamanian
- race\_ethnicity\_Hawaiian
- race\_ethnicity\_Hmong
- race\_ethnicity\_Japanese
- race\_ethnicity\_Kampuchean
- race\_ethnicity\_Korean
- race\_ethnicity\_Laotian
- race\_ethnicity\_Melanesian
- race\_ethnicity\_Micronesian
- race\_ethnicity\_New Guinean
- race\_ethnicity\_Other
- race\_ethnicity\_Other Asian
- race\_ethnicity\_Pacific Islander
- race\_ethnicity\_Pakistani
- race\_ethnicity\_Polynesian
- race\_ethnicity\_Samoan
- race\_ethnicity\_Thai
- race\_ethnicity\_Tongan
- race\_ethnicity\_Unknown
- race\_ethnicity\_Vietnamese
- race\_ethnicity\_White

- seer\_historic\_stage\_a\_Distant
- seer\_historic\_stage\_a\_In situ
- seer\_historic\_stage\_a\_Localized
- seer\_historic\_stage\_a\_Regional
- seer\_historic\_stage\_a\_Unstaged
- sex\_Female
- spanish\_hispanic\_origin\_Cuban
- spanish\_hispanic\_origin\_Dominican Republic
- spanish\_hispanic\_origin\_Mexican
- spanish\_hispanic\_origin\_Non-Spanish/Non-hispanic
- spanish\_hispanic\_origin\_Other specified Spanish/Hispanic origin (excludes Dominican Republic)
- spanish\_hispanic\_origin\_Puerto Rican
- spanish\_hispanic\_origin\_South or Central American (except Brazil)
- spanish\_hispanic\_origin\_Spanish surname only
- spanish\_hispanic\_origin\_Spanish, NOS; Hispanic, NOS; Latino, NOS
- spanish\_hispanic\_origin\_Uknown whether Spanish/Hispanic or not
- year\_of\_birth
- year\_of\_diagnosis
- month
- newtarget

### A.3 Breast Cancer Feature Selection

The feature set used as input into both the Random Forest and Neural Network models, after the transformation described in Section is given below and also available in full detail in the file `NewPatientBreastML.html`.

- cs\_tumor\_size
- elevation
- grade\_moderately differentiated
- grade\_poorly differentiated
- grade\_ndifferentiated; anaplastic
- grade\_well differentiated
- histology\_recode\_broad\_groupings\_adenomas and adenocarcinomas
- histology\_recode\_broad\_groupings\_adnexal and skin appendage neoplasms
- histology\_recode\_broad\_groupings\_basal cell neoplasms
- histology\_recode\_broad\_groupings\_complex epithelial neoplasms
- histology\_recode\_broad\_groupings\_cystic, mucinous and serous neoplasms
- histology\_recode\_broad\_groupings\_ductal and lobular neoplasms
- histology\_recode\_broad\_groupings\_epithelial neoplasms, NOS
- histology\_recode\_broad\_groupings\_nerve sheath tumors
- histology\_recode\_broad\_groupings\_unspecified neoplasms
- lat

- laterality\_Bilateral involvement, lateral origin unknown; stated to be single primary
- laterality\_Paired site, but no information concerning laterality; midline tumor
- laterality\_Right: origin of primary
- lng
- marital\_stats\_at\_dx\_Divorced
- marital\_stats\_at\_dx\_Married (inclding common law)
- marital\_stats\_at\_dx\_Separated
- marital\_stats\_at\_dx\_Single (never married)
- marital\_stats\_at\_dx\_Unknown
- marital\_stats\_at\_dx\_Unmarried or domestic partner
- marital\_stats\_at\_dx\_Widowed
- month\_of\_diagnosis\_Apr
- month\_of\_diagnosis\_Aug
- month\_of\_diagnosis\_Dec
- month\_of\_diagnosis\_Feb
- month\_of\_diagnosis\_Jan
- month\_of\_diagnosis\_Jul
- month\_of\_diagnosis\_Jun
- month\_of\_diagnosis\_Mar
- month\_of\_diagnosis\_May
- month\_of\_diagnosis\_Nov
- month\_of\_diagnosis\_Oct
- month\_of\_diagnosis\_Sep
- race\_ethnicity\_Amerian Indian, Aletian, Alaskan Native or Eskimo
- race\_ethnicity\_Asian Indian
- race\_ethnicity\_Black
- race\_ethnicity\_Chinese
- race\_ethnicity\_Japanese
- race\_ethnicity\_Melanesian
- race\_ethnicity\_Other
- race\_ethnicity\_Other Asian
- race\_ethnicity\_Pacific Islander
- race\_ethnicity\_Thai
- race\_ethnicity\_Unknown
- race\_ethnicity\_Vietnamese
- race\_ethnicity\_White
- seer\_historic\_stage\_a\_Distant
- seer\_historic\_stage\_a\_In sit
- seer\_historic\_stage\_a\_Localized
- seer\_historic\_stage\_a\_Unstaged
- sex\_Female
- spanish\_hispanic\_origin\_Cuban
- spanish\_hispanic\_origin\_Mexican

- spanish\_hispanic\_origin\_Non-Spanish/Non-hispanic
- spanish\_hispanic\_origin\_Other specified Spanish/Hispanic origin (excludes Dominican Republic)
- spanish\_hispanic\_origin\_Spanish surname only
- spanish\_hispanic\_origin\_Spanish, NOS; Hispanic, NOS; Latino, NOS
- year\_of\_birth
- year\_of\_diagnosis
- month
- newtarget

#### A.4 Pseudocode for the Data Transformation

```
def train(X, T, D)
    // X, T, D are the original dataset
    X' = []
    D' = []

    // the transformation
    for each index i in X:
        for t=1 to T[i]:
            new_D = (0 if t < T[i], else D[i])
            append new_D to D'
            new_X = (X[i], t)
            append new_X to X'

    return a decision tree trained on (X', D')
```

```
def pmf(h, X)
    // X is a single datapoint
    // returns an array A where A[i] = P(Y = i | X)
    A = []
    p_so_far = 1 // this is p(T >= t | X)
    for t = 1 to (the last month where h has any data):
        // h knows p(T = t | T >= t, X), we call this p_cur
        p_cur = h's prediction for (X, t)
        append (p_so_far * p_cur) to A
        p_so_far *= (1 - p_cur)
```

## B Model Architecture and Python Code

## C GitHub Repositories

Please always give a title also for appendices.

## Acknowledgments

This is the most common positions for acknowledgments. A macro is available to maintain the same layout and spelling of the heading.

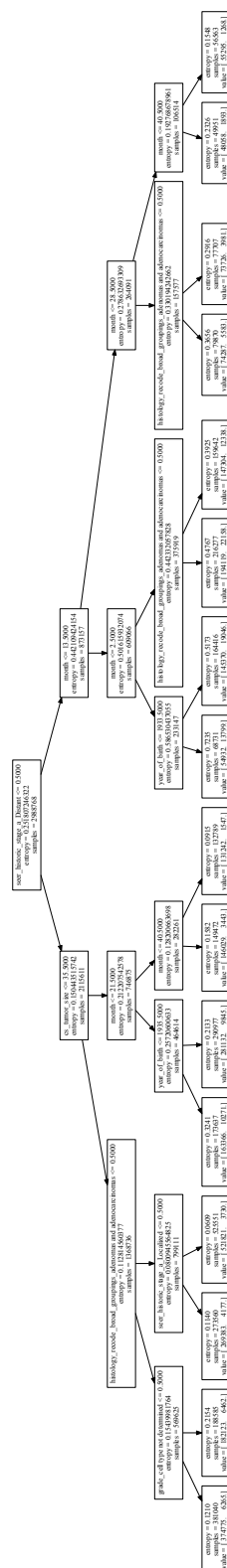
**Note added.** This is also a good position for notes added after the paper has been written.

## References

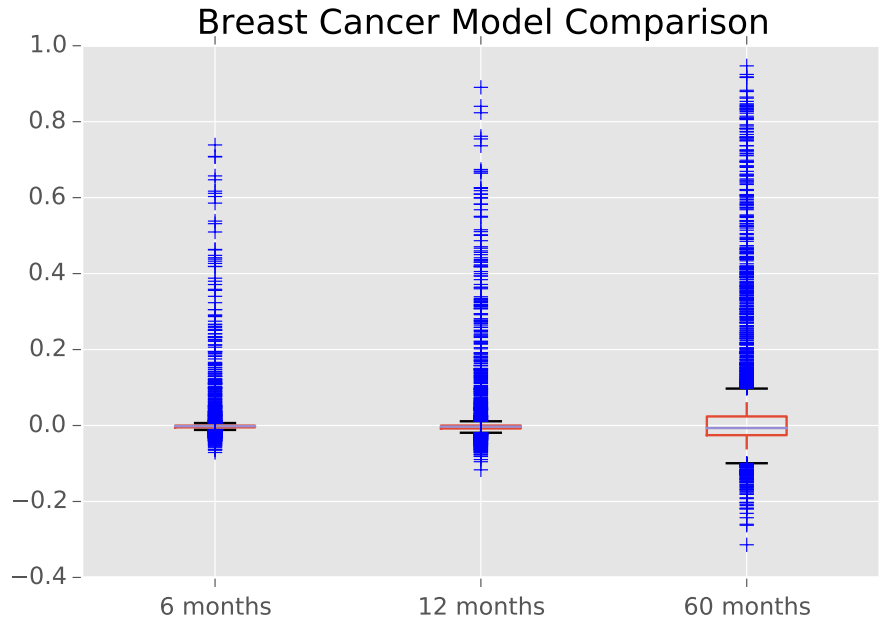
- [1] Ben Kuhn. Decision trees for survival analysis. [http://www.benkuhn.net/survival-trees,year="2016](http://www.benkuhn.net/survival-trees,year=2016) (accessed 14 Jan 2016)",.
- [2] Michael Bowles. *Machine Learning in Python: Essential Techniques for Predictive Analysis*. Wiley, 2015.
- [3] United States Census Bureau. 2010 fips code files for counties - geography - u.s. census bureau. <https://www.census.gov/geo/reference/codes/cou.html>, 2016 (accessed 18 Jan 2016).
- [4] Cam Davidson-Pilon. Quickstart – lifelines 0.8.0.1 documentation. <http://lifelines.readthedocs.org/en/latest/Quickstart.html>, 2016 (accessed 14 Jan 2016).
- [5] Google Developers. The google maps elevation api | google maps elevation api | google developers. <https://developers.google.com/maps/documentation/elevation/intro?hl=en>, 2016 (accessed 18 Jan 2016).
- [6] Google Developers. The google maps geocoding api | google maps geocoding api | google developers. <https://developers.google.com/maps/documentation/geocoding/intro>, 2016 (accessed 18 Jan 2016).
- [7] Allen Downey. *Think Stats*. O'Reilly Media, 2014.
- [8] IOBS. Supplemental material | paperdata. <https://github.com/doolingdavid/PAPERDATA.git>, 2016 (accessed 18 Jan 2016).
- [9] Rabbert Klein. Black holes and their relation to hiding eggs. *Theoretical Easter Physics*, 2010. (to appear).
- [10] Johann A. Makowsky, Saharon Shelah, and Jonathan Stavi.  $\Delta$ -logics and generalized quantifiers. *Annals of Mathematical Logic*, 10:155–192, 1976.
- [11] National Cancer Institute, the Surveillance, Epidemiology, and End Results Program. About the SEER Program - SEER. <http://seer.cancer.gov/about>, 2016 (accessed 14 Jan 2016).
- [12] National Cancer Institute, the Surveillance, Epidemiology, and End Results Program. Documentation for ASCII Text Data Files - SEER Datasets. <http://seer.cancer.gov/data/documentation.html>, 2016 (accessed 15 Jan 2016).
- [13] Sebastian Raschka. *Python Machine Learning Essentials*. Packt Publishing, 2015.
- [14] Hyunjung Shin and Yonghyun Nam. A coupling approach of a predictor and a descriptor for breast cancer prognosis. *BMC MEDICAL GENOMICS*, 7(1), MAY 8 2014. 3rd Annual Translational Bioinformatics Conference (TBC) / ISCB-Asia, Seoul, SOUTH KOREA, OCT 02-04, 2013.



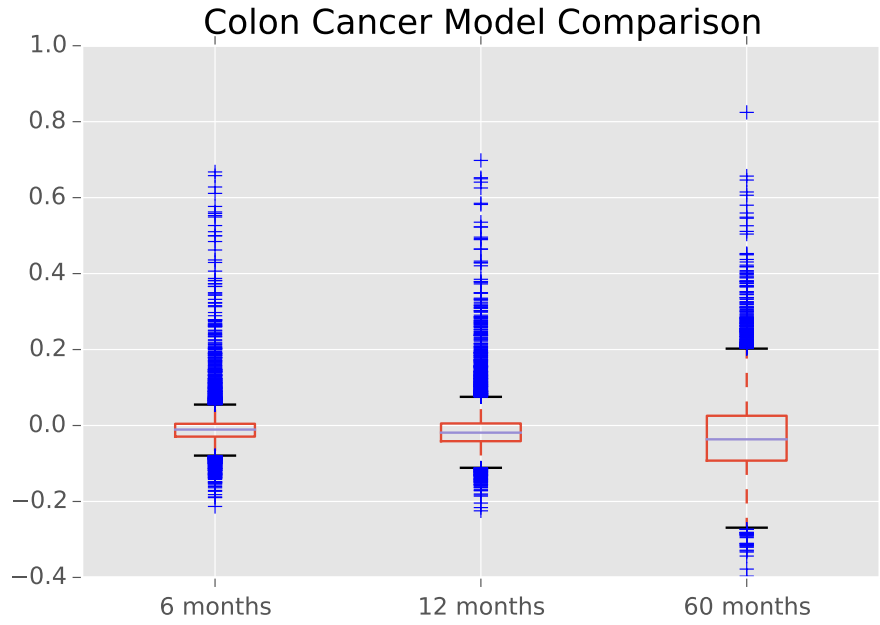
- [15] Hamed Majidi Zolbanin, Dursun Delen, and Amir Hassan Zadeh. Predicting overall survivability in comorbidity of cancers: A data mining approach. *DECISION SUPPORT SYSTEMS*, 74:150–161, JUN 2015.



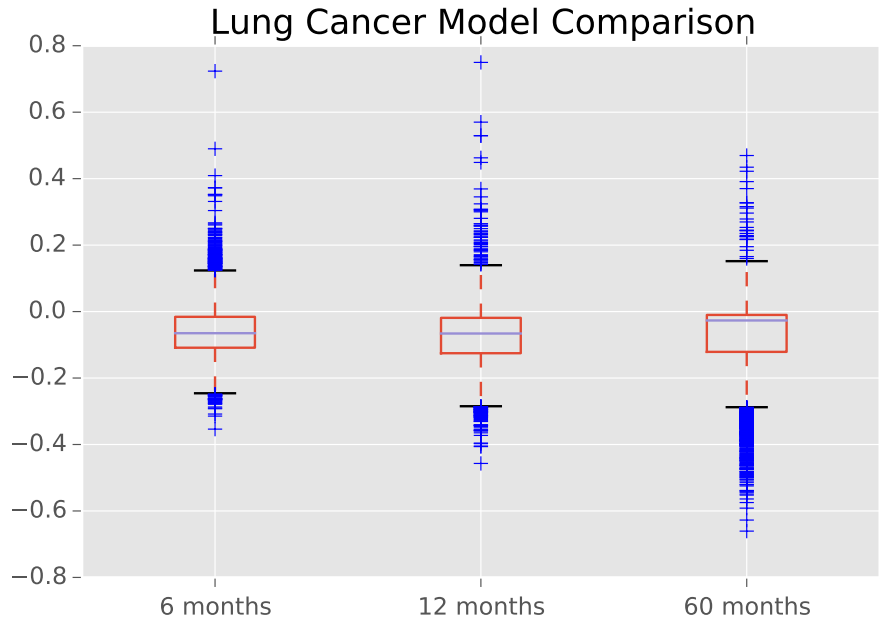
**Figure 4.** The top levels of a decision tree trained on the Lung Cancer training data.



**Figure 5.** Box plots showing the distributions of the signed difference between the MLP model's prediction for the probability of surviving 6 months and the Random Forest model's prediction of the same quantity for breast cancer. The plot shows the same quantity for the 12 and 60 months classifiers. It is apparent from the figures that the outliers are due to the neural network models predicting higher survival probabilities than the random forest for some few cases. These differences were evaluated for the 3300 test patients in the breast cancer data.



**Figure 6.** Box plots showing the distributions of the signed difference between the MLP model’s prediction for the probability of surviving 6 months and the Random Forest model’s prediction of the same quantity for colon cancer. The plot shows the same quantity for the 12 and 60 months classifiers. It is apparent from the figures that the outliers are due to the neural network models predicting higher survival probabilities than the random forest for some few cases. These differences were evaluated for the 5654 test patients in the colon cancer data.



**Figure 7.** Box plots showing the distributions of the signed difference between the MLP model's prediction for the probability of surviving 6 months and the Random Forest model's prediction of the same quantity for lung cancer. The plot shows the same quantity for the 12 and 60 months classifiers. These differences were evaluated for the 5654 test patients in the colon cancer data. The Interquartile Ranges for lung cancer are visibly larger than those for breast cancer and colon cancer shown in fig 5 and fig 6.