Chapter

# An Introduction to Survival Analytics, Types, and Its Applications

*Sheik Abdullah Abbas, Selvakumar Subramanian, Parkavi Ravi, Suganya Ramamoorthy and Venkatesh Munikrishnan*

## Abstract

In today's world, data analytics has become the integral part of every domain such as IOT, security, healthcare, parallel systems, and so on. The importance of data analytics lies at the neck of what type of analytics to be applied for which integral part of the data. Depending upon the nature and type of data, the utilization of the analytical types may also vary. The most important type of analytics which has been predominantly used up in health-care sector is survival analytics. The term survival analytics has originated from a medical domain of context which in turn determines and estimates the survival rate of patients. Among all the types of data analytics, survival analytics is the one which entirely depends upon the time and occurrence of the event. This chapter deals with the need for survival data analytics with an explanatory part concerning the tools and techniques that focus toward survival analytics. Also the impact of survival analytics with the real world problem has been depicted as a case study.

**Keywords:** classification, data analytics, statistics, survival analytics, prediction, parametric models

## 1. Introduction to survival analytics

Survival analysis refers to a branch of statistical analysis domain that evaluates the effect of predictors on *time until an event*, rather than the *probability of an event*, occurs. It is used to analyze data in which the time until the event is of interest. As the name indicates, this method has origins in the field of medical research for evaluating the impact of medicines or medical treatment on time until death. Survival analysis is also known as reliability analysis in the engineering discipline, duration analysis in the economics discipline, and event history analysis in the sociology discipline.

The term is originated from a medical context in which it has been used to estimate the survival rate of patients. Data classification can be dealt explicitly with the process and paradigms available in survival analytical models [1]. The process of survival analytics can be explored through various techniques such as:

- Life tables
- Kaplan-Meier analysis
- Survivor and hazard function rates
- Cox proportional hazards regression analysis
- Parametric survival analytic models
- Survival trees
- Survival random forest

## 2. Metrics for measurement in developing survival models

The process of survival analytics mainly depends on time and occurrence of the event. In survival analytics, time-varying covariates are the variables considered, which change with accordance to the occurrence of the event [2]. The process of survival analytics can be signified and measured using the following measurements:

1. Event time distribution

The event time distribution corresponding for an event to occur with respect to the time function $t$ is defined in Eq. (1) as

$$f(t) = \lim_{\Delta t \to 0} \frac{p(t \le T < t + \Delta t)}{\Delta t} \tag{1}$$

where $t$ is the occurrence of an event at a time t and $\Delta t$ is the small change in time t with accordance to the event.

The distribution varies with accordance to a small change in time for the event that tends to happen for the given function $f(t)$.

2. Cumulative event time distribution

The cumulative event time distribution for the given function $f(t)$ is defined in Eq. (2) as

$$F(t) = P(T \le t) = \int_0^t f(u)du \tag{2}$$

The time period is estimated to be from 0 to $t$.

3. Survival function

The survival function provides the probability estimate in which the corresponding object of life will have existence beyond the period of time $t$. This measure is also termed to be the survivor function or reliability function. The measure can be estimated using the following Eq. (3):

$$S(t) = 1 - F(t) = P(T > t) = \int_0^\infty f(u)du \tag{3}$$

For the condition $S(0) = 1$ and $S(\infty) = 0$, the following relationship holds:

$$f(t) = \frac{dS(t)}{dt} \tag{4}$$

4. Hazard function

The hazard function is also termed to be the hazard rate or the value of mortality, which is then the ratio among the probability density function and the survivor function which is depicted in Eq. (5) as

$$h(t) = \frac{f(t)}{S(t)} \tag{5}$$

where $h(t)$ is the hazard function, $f(t)$ is the probability density function, and $S(t)$ is the survival function.

## 3. Model classification in survival analytics

The process behind survival analytics is different when compared to predictive and descriptive analytics [3]. Here, the time component is an important factor which efficiently determines the success or failure of a model. The following **Figure 1** illustrates the model to be classified under survival analytics [4]. Different sorts of functions are adaptable with different models based on the metric to be used with time as a component for the event to occur. The main target is to determine the right model to be chosen for the observed survival analytic data. In parametric model analysis, the survival curve depends only on the shape of the model with its function value.

The shape of the model can be estimated with regard to the characteristics of a nonparametric model. As an outcome, the shape of the hazard function also varies with regard to time. Some of the examples corresponding to hazard shapes are:

- Increasing hazard
- Decreasing hazard
- Constant hazard
- Convex bathtub-shaped hazard

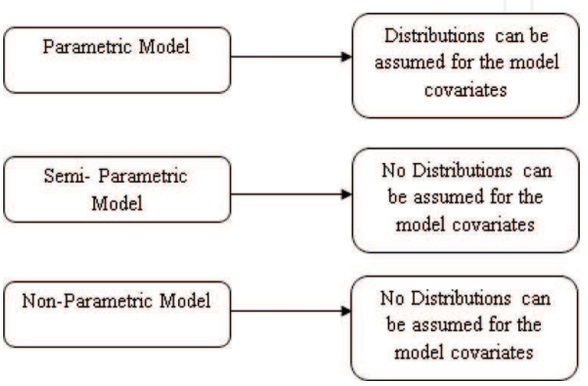


**Figure 1.**
*Model classification in survival analytics.*

Parametric survival analysis can be understood with the available forms of distributions. The distributions usually convey the efficacy in terms of probability curve with time analysis. The following are the distributions used to estimate the output measured by the survival curve:

1. Normal distribution
2. Weibull distribution
3. Exponential distribution
4. Lognormal distribution
5. Gamma distribution
6. Uniform distribution
7. Log-logistic distribution

The following illustrations provide an overview with regard to each of the distributions in detail:

## 4. Parametric survival analytical models

### 4.1 Exponential distribution

The exponential distribution is also known as negative exponential distribution. Exponential distribution is defined as a process in which events occur continuously and independently at a constant average rate. The exponential distribution is defined as

$$f(t) = \lambda e^{-\lambda t} \tag{6}$$

The survivor function is then estimated as

$$S(t) = \lambda e^{-\lambda t} \tag{7}$$

The hazard rate is then estimated as

$$h(t) = \frac{f(t)}{S(t)} \tag{8}$$

$$h(t) = \lambda \tag{9}$$

Hence, from Eq. (9), it should be noted that the hazard rate is independent of time and therefore the risk corresponding to the event remains to be same. The following **Figures 2** and **3** illustrate the event time with respect to the hazard rate.

### 4.2 Weibull distribution

The Weibull distribution is defined as a continuous probability distribution. The expression is defined in Eq. (10) as

$$f(t) = k\rho\,(\rho t)^{k-1} \exp\left[-(\rho t)^{k}\right] \tag{10}$$

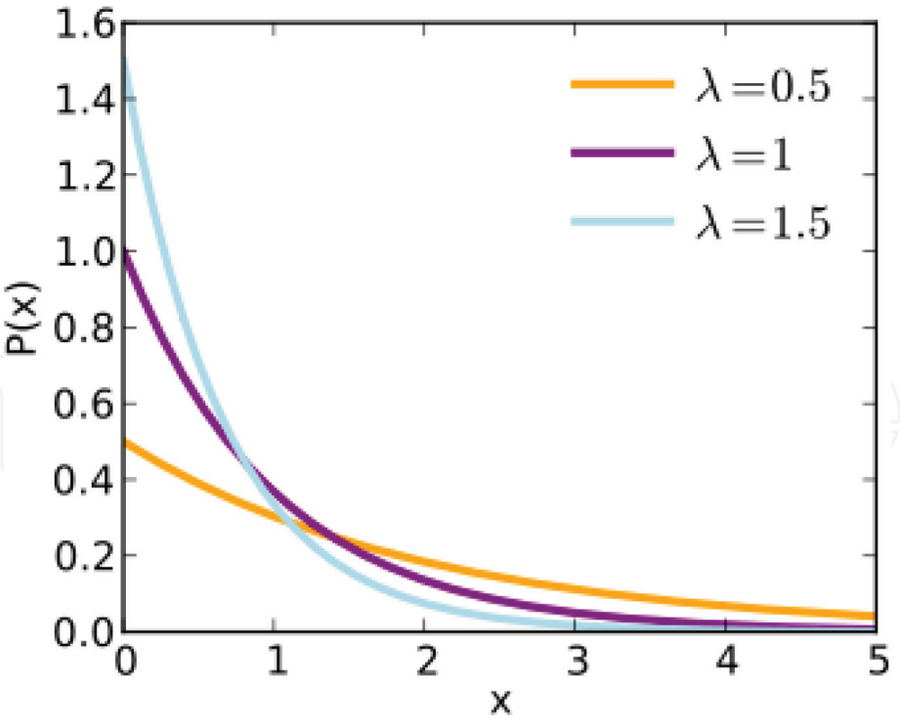


**Figure 2.**
*Probability density function.*

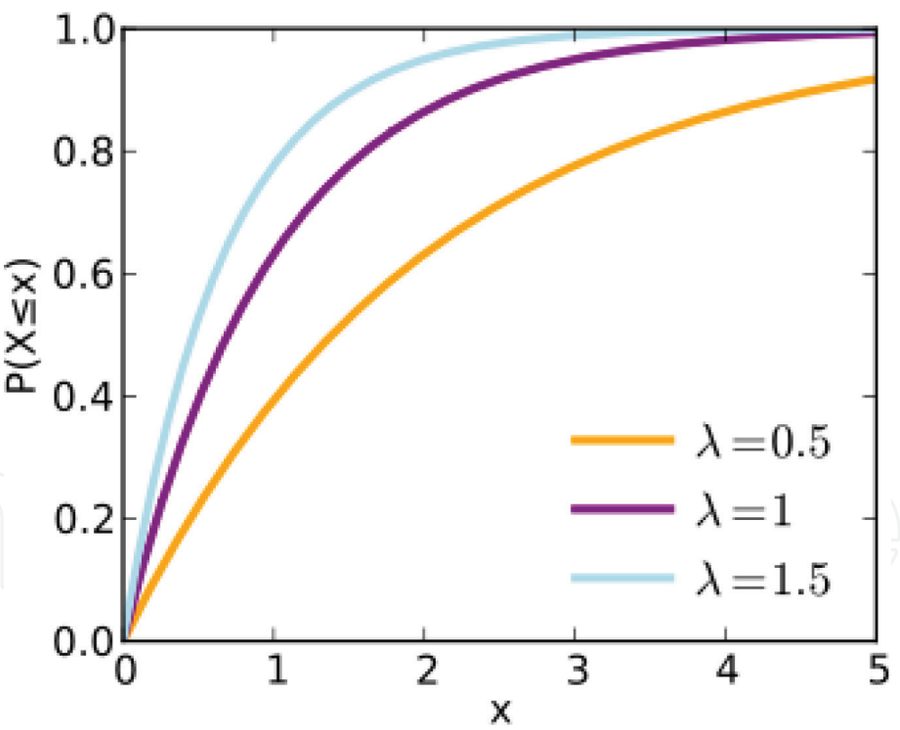


**Figure 3.**
*Cumulative distribution function.*

The evaluation of hazard rate is given as

$$h(t) = k\rho(\rho t)^{k-1} \tag{11}$$

Hence, for this case, the hazard rate depends on time which can be in either increasing or decreasing mode. The following **Figures 4** and **5** depict the value of hazard rate with respect to time $t$.

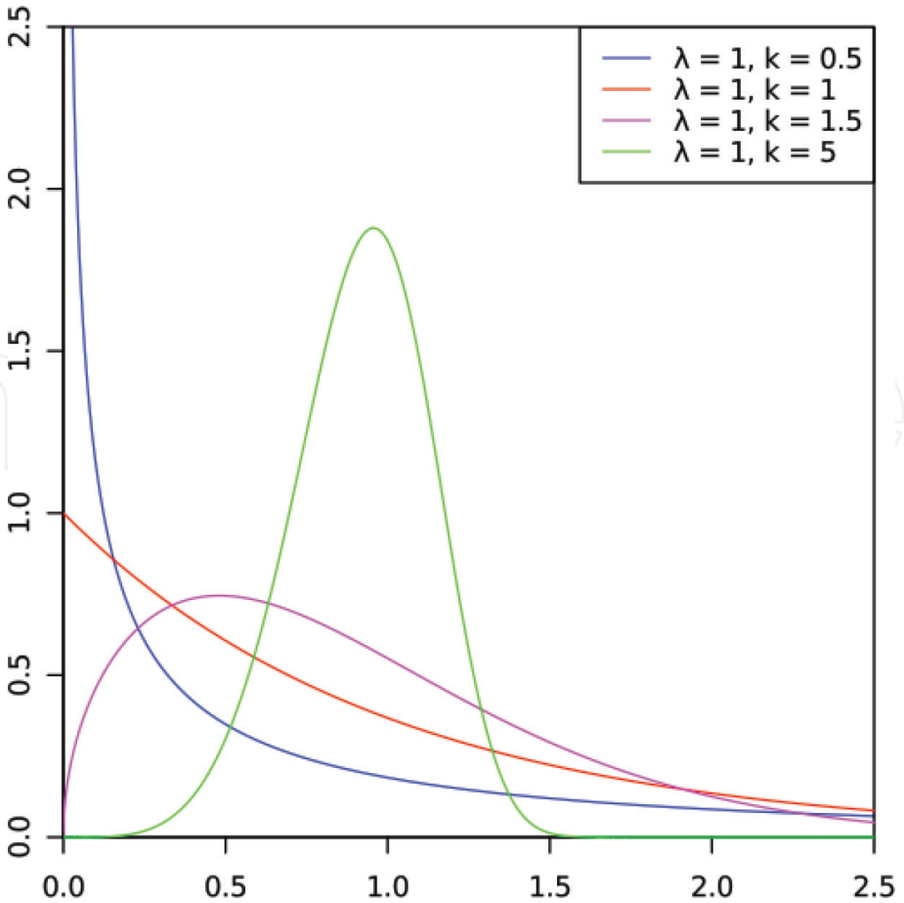


**Figure 4.**
*Probability distribution function.*

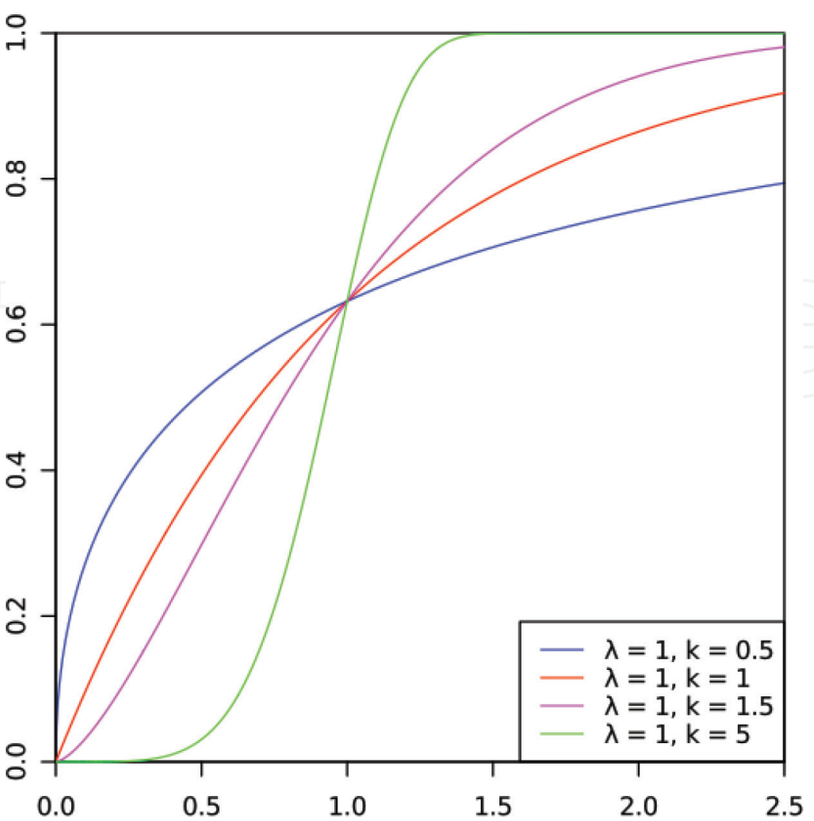


**Figure 5.**
*Cumulative distribution function.*

### 4.3 Log-logistic distribution

The log-logistic distribution is defined as a continuous probability distribution with negative random variable. In economics discipline log-logistic distribution is also known as the Fisk distribution in economics. Log-logistic is a continuous probability distribution for a nonnegative random variable. The following **Figures 6** and 7 depict the distribution of log-logistic model.

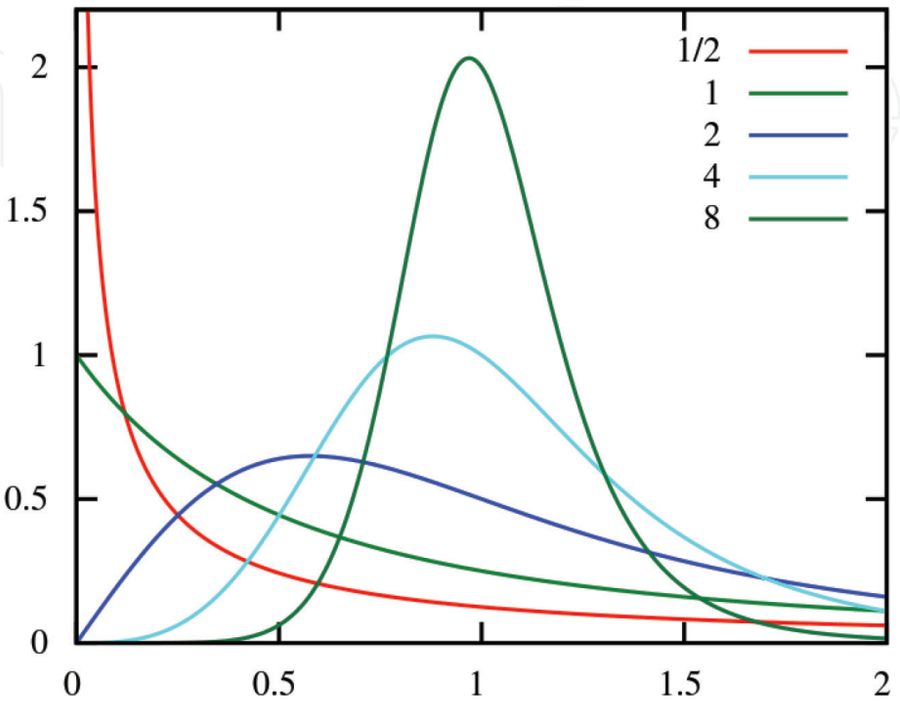


**Figure 6.**
*Probability density function.*

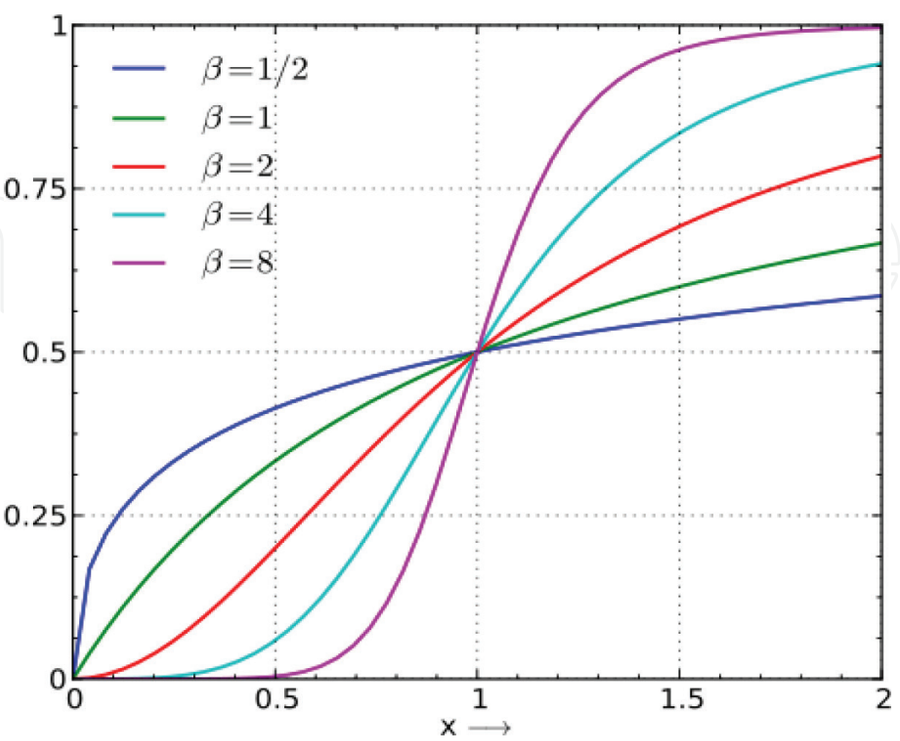


**Figure 7.**
*Cumulative distribution function.*

## 5. Kaplan-Meier analytics and Cox regression model

The Kaplan-Meier test is broadly used within the pharmaceutical industry for specifying the expiry data for clinical drug in health-care sector, monitoring the effects of drugs and their gestures on recovery time or critical time. The Kaplan-Meier test is a statistical method that really works well for effective cancer treatments. This test determines the patient's survival time between two groups. For clinical drug examination, a successful test indicates that the group of people taking the new drug has a shorter time to improvement or death than the group of people taking a place [5].

If the value of censoring is not available, then the value of the KM estimator for $S(t)$ is found to be the same proportional value with respect to $t$. If the value of censoring is present, then the following steps are followed:

Step 1: Order the event times in ascending order of levels $t_1 < t_2 < t_3 < \cdots < t_k$.

Step 2: At each time $t_j$, there are about $n_j$ individuals who are subjected to risk of the event.

Step 3: Let $d_j$ be the number of individuals who die at $t_j$ (churn, respond).

Therefore, the KM estimator is defined as

$$S(t) = \prod_{J:t_1 \le t} \left(1 - \frac{d_j}{n_j}\right) = S(t-1) \bullet \left(1 - \frac{d_t}{n_t}\right) \tag{12}$$

If there exists uniqueness in event times, then the KM estimator is measured using the life table for grouping of event times as expressed as

$$S(t) = \prod_{J:t_1 \le t} \left(1 - \frac{d_j}{n_j - c_j/2}\right) \tag{13}$$

where

$n_j$ is the number of individuals at risk.

$d_j$ is individuals who die at a specified time.

The implementation of KM estimator can be extended by applying statistical tests such as hypothesis testing, Wilcoxon test, and likelihood ratio test. With exploratory data analytics along with KM estimator, the patterns and insights can be determined more efficiently from the data.

The second popular survival analysis method used for prediction is Cox. It is also known as the Cox model but often referred to as Cox regression. It is more popular in the Web of Science. More than 38,000 articles are cited indexing the Cox regression method.

There are some other statistical/analytical methods available that can predict time until an event, but survival analysis methods have the unique feature of considering the past history/experiences. Although these latter cases do not have a date for the target event, they are an integral part of the analysis. The terminology used in survival analysis is called censored cases.

Another formal definition for survival analysis is, it is basically defined as a set of methods for analyzing data where the outcome variable is the time/instance until the occurrence of an event of interest. The event can be an uncertainty accident, death, occurrence of a disease, or planned ones—marriage, divorce, etc. The time to event or survival time can be measured in various scales of time periods (days, weeks, years, etc.).

For example, if the event of interest is mild heart attack, then the survival time can be the time in years until a person develops a heart attack. Choose any survival methods that are discussed above. In survival analysis, time is a primary factor.

The advantage of Cox regression over Kaplan-Meier is that it can accommodate any number of predictors, i.e., chances of getting heart attack, rather than group membership only. As is the case for all regression methods, there are two potential benefits of analysis using Cox regression: *predictor ranking*, with each predictor's effect measured greater than the predicator's threshold effect or less than the predicator's threshold effect and the ability to *make predictions* with the regression results. Predictor rankings facilitate the analyst to recognize the factors that have the most influence on time to an event, and the regression results can be used to estimate the amount of until an event for a specific profile of any subject [6].

### 5.1 Different types of censoring

Data can be either right, left, or interval censored. It is the sum of defined time $t_o$, and the event of interest takes place at $t_o + t$, where t is an unknown factor and the event is only known to have occurred at $t_o + c$ and the data is censored with a censored time, c.

***Right censoring*** is the most common, occurring when the true event time is greater than the censored time, when $c < t$. It often arises when the event of interest has not occurred by the end of study and the subject has been lost to follow-up.

***Left censoring*** is the opposite, occurring when the true event time is less than the censored time, when $c > t$.

***Interval censoring*** is a concatenation of the left and right censoring, when the time is known to have occurred between two time points: $c_1 < t < c_2$.

Censoring is an important matter in survival analysis, signifying a particular type of missing data. Censoring is a random and non-informative study, and it is usually required in order to avoid bias in a survival analysis The interpretation of Cox regression and Kaplan results depends two factors: positive (e.g., a sale) or negative (e.g., product failure).

## 6. Case study for churn prediction

The following graphical illustrations depict the implementation of churn prediction and model deployment using RapidMiner. The algorithm used for analysis is decision tree [7]. The implementation has been done with the lift chart analysis with evaluation in performance metrics. The attributes in the dataset includes person ID, churn status, gender, age, region code, transaction count, average balance, and total accounts.

The case study majorly explores with an application that is most probably used up with churn analysis. Nowadays, churn prediction is majorly analyzed in most of the industries to track the historical learning with the customers. The entire customer demographic data is analyzed day to day with regard to the maintenance of business relationships, customer transactions, products purchased, and the survey that has been obtained with regards to the business attractions. To make an exploration in this application, we have used up RapidMiner tool for the entire survival rate estimation and analysis of customers in an organization. The above **Figure 8** provides the selection of application with regard to churn prediction for estimating the survival rate of customers.

RapidMiner is one of the good statistical and analytical tools which is mostly practiced in industries and academic institutions. Rapid miner provides a good insight for statisticians and mathematical experts to observe the insights and patterns that lie within the given data. The following **Figure 9** explores the analytical results observed with RapidMiner.

All the incorporations in RapidMiner are made through the process connection through wires. The workflow of each process is written through Java. The process diagram depicts the step-by-step flow of algorithmic model development through drag option. **Figure 10** provides a complete overview with regard to the process creation for churn prediction analysis. The algorithm used for the development of the model is decision tree classification algorithm [8]. Decision tree algorithm provides a tree-like structure in a top-down fashion with a single root node and a number of

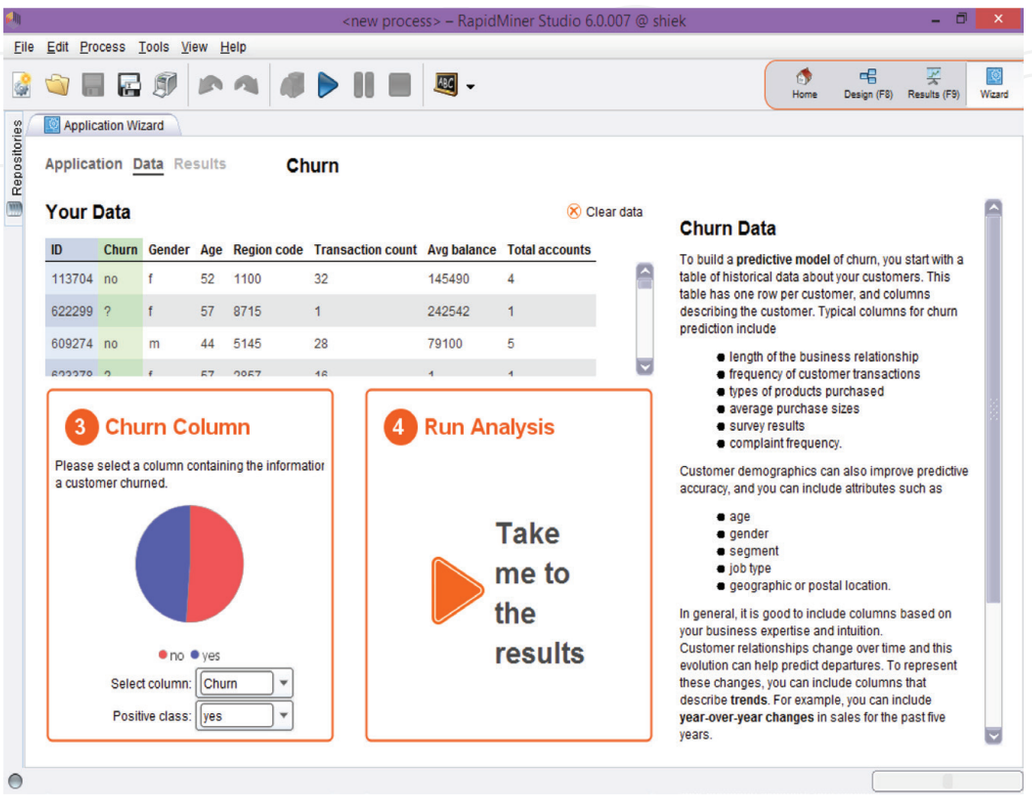


**Figure 8.**
*Data selection for churn analysis.*

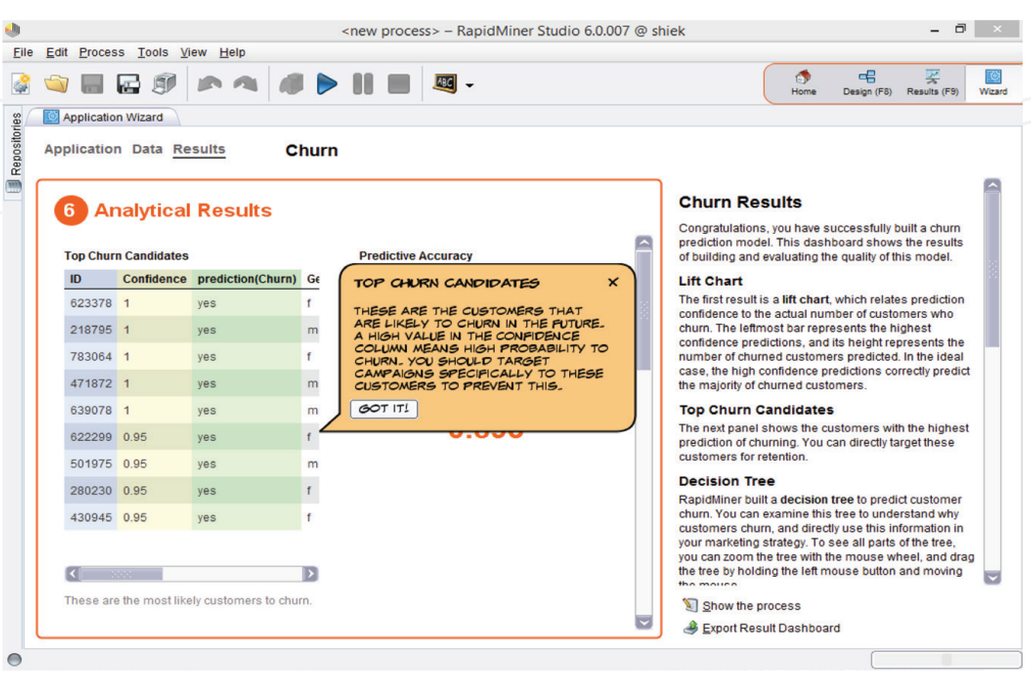


**Figure 9.**
*Analytical results observed for churn model.*

leaf nodes with a terminating condition. The working of the algorithm depends on the splitting criterion to be used up for analysis (**Figures 11** and **12**).

Lift chart shows the effectiveness of the predictive model in which it has to be developed. It generally provides the ratio between the predicted values to that of the actual one. In **Figure 13** for churn analysis, the chart provides the ratio between the confidence value and the count observed for churn analysis. Thus, churn prediction is employed for tracking the survival rate of customers with survival analytics. Survival analytics model can be deployed more efficiently for tracking the rate of patients in medical domain. The realm of health informatics lies at the heart of existence of subjects concerned with specific disease. The existence and

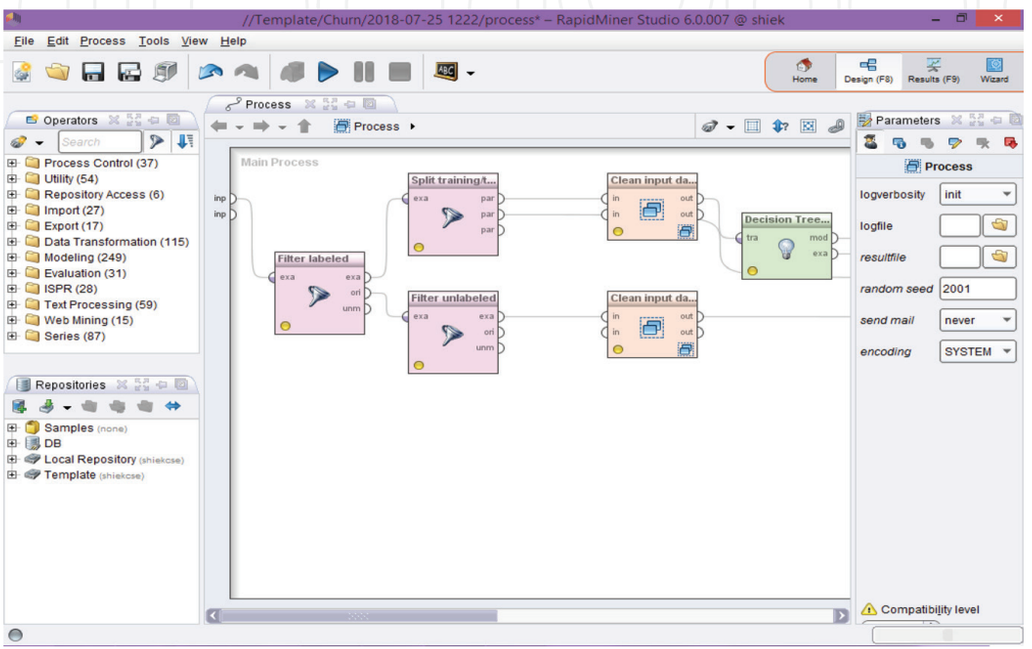


**Figure 10.**
*Process diagram in a design view.*

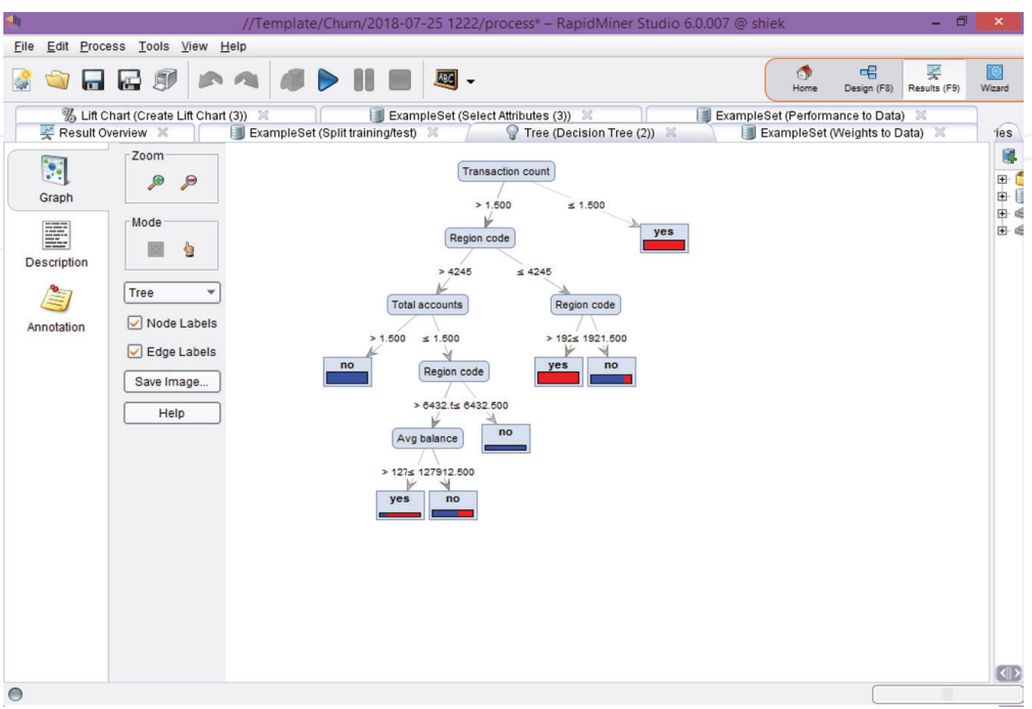


**Figure 11.**
*Generated tree with decision tree algorithm.*

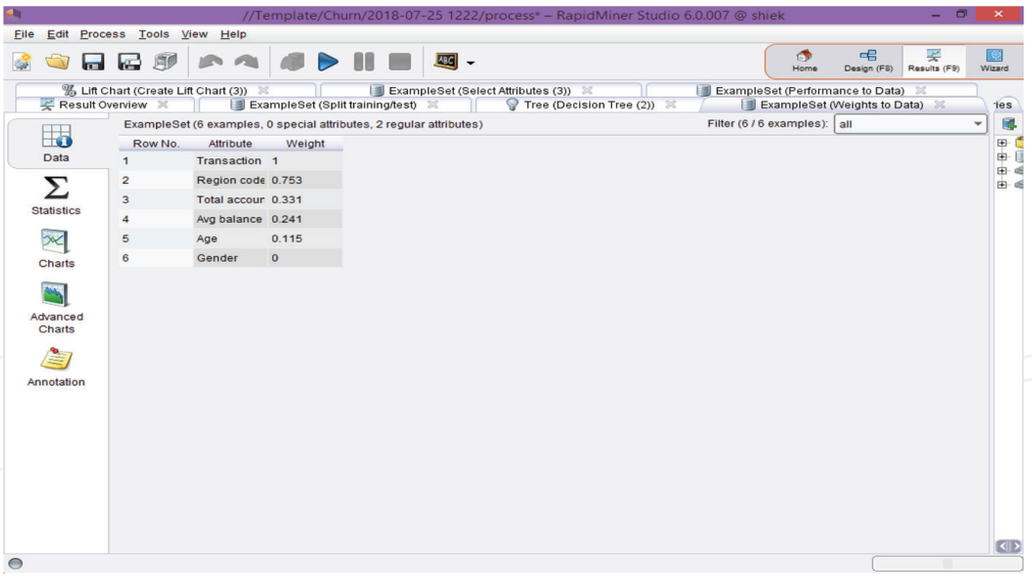


**Figure 12.**
*Performance evaluation.*

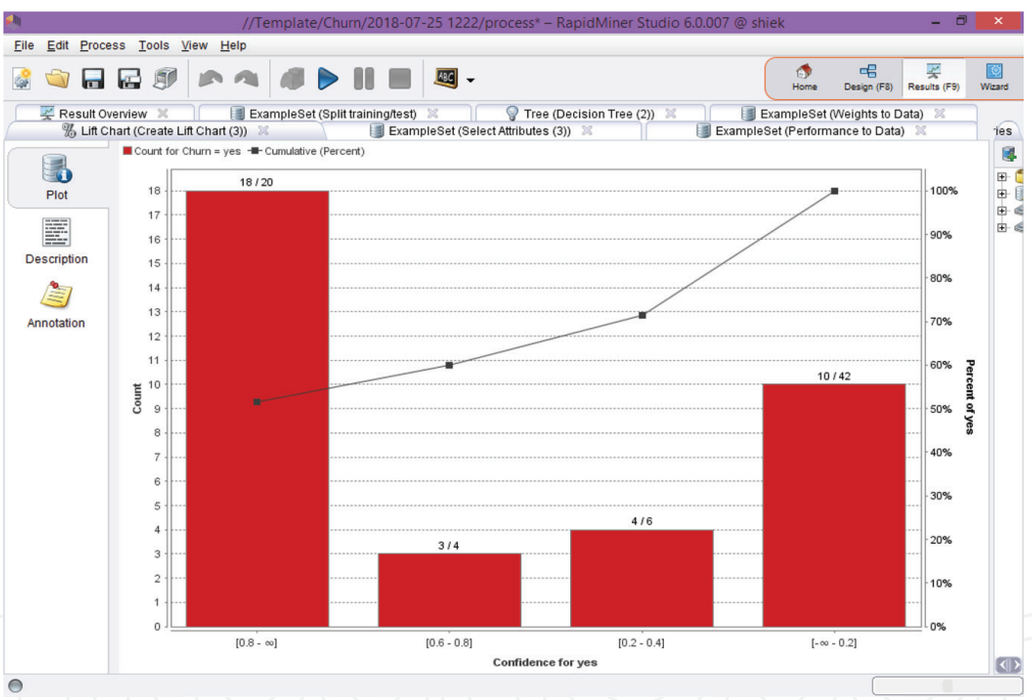


**Figure 13.**
*Lift chart for churn analysis.*

the nonavailability of subjects with regard to the specific disease can be learnt with patterns and explorations through survival analytics models.

## 7. Case study using Kaplan-Meier analytics

Consider there are about 200 subjects of patient's records which have been tracked over a period of time. The tracking is made in such a way that the total number of patients has to be confirmed with first year, second year, and third year, and so on. If all the subjects have existed for the given duration, then there will not be a case for probability of occurrence with regard to each of the subjects. To illustrate this complicated situation, consider the following scenario in **Table 1**.

**Condition**:

1. Out of 200 subjects, 6 became unavailable and 10 have been found to be dead at the end of the first year.

2. With the remaining subjects, 6 became unavailable and 20 have been found to be dead at the end of the second year.

3. With the remaining subjects, 6 became unavailable and 30 have been found to be dead at the end of third year.

4. With the remaining subjects, 6 became unavailable and 40 have been found to be dead at the end of fourth year.

5. With the remaining subjects, 6 became unavailable and 50 have been found to be dead at the end of fifth year.

| Time period | At risk | Become unavailable (censored) | Died | Survived |
|---|---|---|---|---|
| Year 1 | 200 | 6 | 10 | ? |
| Year 2 | ? | 6 | 20 | ? |
| Year 3 | ? | 6 | 30 | ? |
| Year 4 | ? | 6 | 40 | ? |
| Year 5 | ? | 6 | 50 | ? |

**Table 1.**
*Scenario of individuals for a specified disease.*

| Time period | At risk | Become unavailable (censored) | Died | Survived |
|---|---|---|---|---|
| Year 1 | 200 | 6 | 10 | 190 |
| Year 2 | 184 | 6 | 20 | 164 |
| Year 3 | 158 | 6 | 30 | 128 |
| Year 4 | 122 | 6 | 40 | 82 |
| Year 5 | 76 | 6 | 50 | 26 |

**Table 2.**
*Observed solution as per Kaplan-Meier condition.*

| Time period | At risk | Become unavailable (censored) | Died | Survived | Kaplan-Meier survival probability estimate |
|---|---|---|---|---|---|
| Year 1 | 200 | 6 | 10 | 190 | (190/200) = 0.95 |
| Year 2 | 184 | 6 | 20 | 164 | (190/200) × (164/184) = 0.84 |
| Year 3 | 158 | 6 | 30 | 128 | (190/200) × (164/184) × (128/158) = 0.70 |
| Year 4 | 122 | 6 | 40 | 82 | (190/200) × (164/184) × (128/158) × (82/122) = 0.46 |
| Year 5 | 76 | 6 | 50 | 26 | (190/200) × (164/184) × (128/158) × (82/122) × (26/76) = 0.15 |

**Table 3.**
*Kaplan-Meier probability estimate.*

For this scenario we can determine the list of individuals who are all became unavailable at the end of the given time period. Use Kaplan-Meier analytics to determine the individuals who are at risk and what would be the probability estimate for the individuals survived at the end of the fifth year.

**Solution**:

Step 1: Kaplan-Meier suggested that the subjects that became unavailable during the given time period can be counted among with those who survive through the end but are removed or deleted from the total number of individuals who are subjected to risk for the next given time period. With these conventions, the formulation is described in **Table 2**.

Hence, from **Table 2**, it has been observed that at the end of fifth year, 26 individuals have survived from the set of 200 individuals who were subjected to a specified disease. The next is to determine the Kaplan-Meier probability estimate for each of the time intervals $t$ with regard to the conditional probability. The following **Table 3** provides the probability estimate for 5 years of risk analysis.

From **Table 3**, it has been observed that at the end of fifth year, the conditional probability estimate was found to be 0.15% of individuals. Hence, from the perspective of survival probabilistic estimate, we can determine the existence rate of individuals for the given time period $t$.

## Author details

Sheik Abdullah Abbas[1*], Selvakumar Subramanian[2], Parkavi Ravi[1], Suganya Ramamoorthy[1] and Venkatesh Munikrishnan[3]

1 Department of Information Technology, Thiagarajar College of Engineering, Madurai, Tamil Nadu, India

2 Department of Computer Science and Engineering, GKM College of Engineering and Technology, Chennai, Tamil Nadu, India

3 Department of General Medicine, Theni Government Medical College and Hospital, Theni, Tamil Nadu, India

*Address all correspondence to: asait@tce.edu

## References

[1] Sheik Abdullah A, Gayathri N, Selvakumar S, Rakesh Kumar S. Identification of the Risk Factors of Type II Diabetic Data Based Support Vector Machine Classifiers upon Varied Kernel Functions. Lecture Notes in Computational Vision and Biomechanics [Internet]. Switzerland: Springer International Publishing; 2018. pp. 496-505. Available from: http://dx.doi.org/10.1007/978-3-319-71767-8_42

[2] Tsujitani M, Baesens B. Survival analysis for personal loan data using generalized additive models. Behaviormetrika [Internet]. Behaviormetric Society of Japan; 2012; **39**(1):9-23. Available from: http://dx.doi.org/10.2333/bhmk.39.9

[3] Sheik Abdullah A, Selvakumar S, Abirami AM. An Introduction to Data Analytics. Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence [Internet]. Hershey, USA: IGI Global; 1-14. Available from: http://dx.doi.org/10.4018/978-1-5225-2031-3.ch001

[4] Cox DR. Regression Models and Life-Tables. Breakthroughs in Statistics [Internet]. Switzerland, New York: Springer; 1992;527-541. Available from: http://dx.doi.org/10.1007/978-1-4612-4380-9_37

[5] Banasik J, Crook JN, Thomas LC. Not if but When will borrowers default. The Journal of the Operational Research Society [Internet]. JSTOR; 1999 Dec;**50**(12):1185. Available from: http://dx.doi.org/10.2307/3010627

[6] Crowder M. Classical Competing Risks. UK: CRC Press, Taylor & Francis Group, an Informa Group company; 2001 May 11. Available from: http://dx.doi.org/10.1201/9781420035902

[7] Abdullah AS, Selvakumar S, Karthikeyan P, Venkatesh M, et al. Comparing the efficacy of decision tree and its variants using medical data. Indian Journal of Science and Technology [Internet]. Indian Society for Education and Environment. 2017 May 1; **10**(18):1-8. Available from: http://dx.doi.org/10.17485/ijst/2017/v10i18/111768

[8] Sheik Abdullah A, Suganya R, Selvakumar S, Rajaram S. Data Classification. Handbook of Research on Advanced Data Mining Techniques and Applications for Business Intelligence [Internet]. Hershey, USA: IGI Global; 34-51. Available from: http://dx.doi.org/10.4018/978-1-5225-2031-3.ch003