Predicting Next Purchase Probabilities using Random Survival Forests

Emmanuel KABUGA (kabuga@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Professor Ian Durbach
University of Cape Town, South Africa

18 May 2017

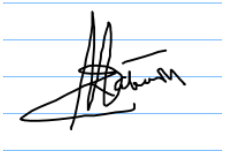
Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

In this project, we examine the problem of identifying (i) what factors are likely to lead a person to buy again, (ii) which customers are most "at risk" of buying again/not buying again. We dealt with the time-to-event data, where the outcome of interest is the purchase occurrence, and there are some explanatory variables that can help in predicting this outcome. A feature of time-to-event data is that it is often right-censored, meaning that we do not observe the event of interest (the purchase) for all individuals in our sample. We use random survival forests (RSF), a relatively new extension of Breiman's random forest (RF) algorithm to deal with right-censored data. In addition to RSF, we run the Cox proportional hazard (PH) model, the gold standard in survival analysis and compare the result. We apply our models to data collected from Australia on product purchases for customers in toothpaste categories.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.



Emmanuel KABUGA, 18 May 2017

Contents

1. Introduction

One of the main goals of marketing is to find out the needs of customers as well as the right moment to satisfy them. Once these features are assessed, they should be used to navigate the marketplace and offer the right products to the right customers at the right moment. Customers are not regularly observed at the marketplace. One of the market questions is " will a customer who has bought a product in the past buy again? And if so, when?" An understanding of this question will enable marketers to keep track of customers and make predictions about gains or losses in market share.

In this project, we consider the following question, which is a simplified but realistic version of one often faced by marketers. Suppose we follow a sample of past customers for a short period of time. For each customer, we record whether they purchase a product or a brand and, if so, when that purchase occurred. In addition to that data, suppose that we have some basic information about a customer's past purchase history (e.g. the number of purchases in the past year, and when the last purchase occurred) and his or her demographics. Our questions are:

(i) how can we model the probability of repurchase as a function of these covariates?

(ii) how accurate is our model?

(iii) can we use the model to inform marketing strategy, by assessing which customers are most "at risk" of buying/not buying?

(iv) what are factors that are likely to influence a customer to buy again?

Our data is a panel data (It displays a subject's behaviour and is derived from observations over a given period. It is also known as longitudinal data) collected from Australia on a sample of 2662 customers on purchases with 35 different brands in toothpaste categories over 16 weeks. This kind of data is often referred to as a "survival data". It is a data that arises when the time from the origin until some event of interest occurs is recorded for each subject. It is characterized by the occurrence of one event at most per subject. A subject who does not experience the event of interest by the end of the follow-up time is said to be censored.

In this project, we try to answer questions aforementioned. We will introduce the random survival forests (RSF) model, an extension of Breiman's random forests (RF) (Breiman, 2001), designed to tackle survival data and fit its model to the data to predict probabilities of next purchase occurrence. In addition to the RSF method, we will run the Cox proportional hazard (PH) model (Cox, 1972) also called the Cox regression model, the gold standard in survival analysis and compare results from both models.

The remainder of our work is organized as follows: In Chapter 2, we introduce the survival analysis and some concepts behind it. In Chapter 3, we highlight the background related to tree-based methods such as regression and classification trees and their corresponding tree-ensemble approaches such as bagging, RF, and RSF. In Chapter 4, we apply our models to the panel data, describe the data, implementations details and compare results from both models. We end up with conclusions in Chapter 5.

2. Survival analysis

2.1 Introduction

Survival analysis is an ensemble of techniques for analyzing survival data. The predictive response is the time till an event of interest occurs. The time-to-event can be expressed in terms of days, weeks, months, years, etc. During the survival analysis process, subjects are followed over a fixed period of time and one is interested in the time of occurrence of an event of interest. Subjects who do not experience the event of interest before the end of fixed time period are said to be right censored. In survival analysis, censoring is a great problem, since it introduces a new kind of data called missing data. Survival approaches involve both censored and uncensored information when predicting. The dependent variable takes into account two main features, the time-to-event and the event status information (occurred or not). Those two pieces of information are known via the survival and hazard functions. In this essay, we apply survival analysis techniques in a business context, where the event of interest is the purchase occurrence. Survival and hazard functions are key features in survival analysis. In this section, those functions are highlighted as well as their estimation.

2.2 Survival function and hazard function

The survival and hazard functions play an important role in summarizing data. Let T be a random variable associated with the survival time t with probability density function $f(t)$. The distribution function associated with T is yielded by

$$F(t) = P(T < t) = \int_0^t f(u)du,$$

it denotes the probability that a customer does not make any purchase at the time less than t.

The survival function $S(t)$ that represents the probability that a customer does not buy a product up to time beyond t, is given by

$$S(t) = P(T \geq t) = 1 - F(t). \tag{2.2.1}$$

At a given time t, the survival function denotes the probability that a customer does not make a purchase up to that time.

Another feature associated with the distribution of T is the hazard function, or instantaneous rate of purchase occurrence, given by

$$h(t) = \lim_{\delta t \to 0} \left[\frac{P(t \leq T < t + \delta t | T \geq t}{\delta t} \right]. \tag{2.2.2}$$

The numerator of this relation gives the conditional probability that a customer will purchase in the interval $[t, t + \delta t)$ given that he has not purchased before. The denominator denotes the width of the interval. Dividing the former by the latter, we get the rate of purchase occurrence per unit of time. As the limit of the width interval tends to zero, we get the instantaneous rate of purchase occurrence. The

numerator can be re-expressed as the ratio of the joint occurrence probability that T is in the interval $[t, t+\delta t)$, over the condition probability $T \geq t$. We have

$$\frac{P(t \leq T < t+\delta t)}{P(T \geq t)},$$

which is the same as

$$\frac{F(t+\delta t) - F(t)}{S(t)},$$

where $F(t)$ is the distribution function of T. It follows that

$$h(t) = \lim_{\delta t \to 0} \left[\frac{F(t+\delta t) - F(t)}{\delta t} \right] \frac{1}{S(t)}.$$

Since

$$\lim_{\delta t \to 0} \left[\frac{F(t+\delta t) - F(t)}{\delta t} \right],$$

denotes the derivative of $F(t)$ respect to t which is $f(t)$, then

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.2.3}$$

The rate of purchase occurrence at time t is given by the density of purchases at that time over the probability of not purchasing up to that time.

using (2.2.1), (2.2.3) becomes

$$h(t) = -\frac{d}{dt}\{\log S(t)\}, \tag{2.2.4}$$

the equation above leads to

$$S(t) = \exp\{-H(t)\}, \tag{2.2.5}$$

where

$$H(t) = \int_0^t h(u)du, \tag{2.2.6}$$

$H(t)$ is a feature commonly used in survival analysis and is either called the **integrated** or **cumulative hazard**.

From (2.2.5),

$$H(t) = -\log S(t). \tag{2.2.7}$$

2.3 Non-parametric procedures

One of the main steps of survival analysis is to display numerically or graphically the summaries of survival times. These summaries are important since they give an idea about how many subjects experience the event of interest or not. Survival data information is nicely summarized via survival and hazard functions using non-parametric approaches. These methods are so called because they do not rely on assumptions. Two methods, the life-table and Kaplan-Meier method, are introduced to estimate the survival and hazard functions.

2.3.1 Estimating the survival function. The survival function evaluates the probability that a given subject will not experience the event of interest.

Life-table estimate of the survival function. One can get the life-table estimate by splitting the period of observation into a series of m time intervals. Let us consider that the $j^{th}\,(j = 1, 2, \ldots, m)$ interval lies between the time t'_j and t'_{j+1}. Let d_j and c_j be respectively the number of customers who bought products and those censored in time period j. Let n_j be the total number of customers. Assuming that censoring happens uniformly all along the j^{th} interval, the average number of customers is

$$n'_j = n_j - \frac{c_j}{2}. \tag{2.3.1}$$

This relation is called the **actuarial assumption**. Throughout the j^{th} interval, the probability of buying a product is estimated by $\frac{dj}{n'_j}$ and the opposite is given by $1 - \frac{d_j}{n'_j}$. Given the k^{th} interval, the probability that a customer does not buy a product up to time beyond t'_k is the product of the probability that a customer does not buy a product along the k^{th} interval as well as throughout $k-1$ previous intervals. The life-table estimate of the survival function is yielded by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n'_j - d_j}{n'_j} \right), \tag{2.3.2}$$

for $t'_k \leq t < t'_{k+1}$, $k = 1, 2, \ldots, m$. The estimated probability is taken to be 1 before the first interval, t'_1 and 0 beyond t'_{m+1}

Kaplan-Meier estimate of the survival function.The estimate of survival function in the k^{th} time interval lying between t_k and t_{k+1} is the probability that one does not make any purchase throughout the k^{th} interval as well as the preceding intervals. It is given by

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{2.3.3}$$

where d_j and n_j represent respectively the number of customers who made purchases and the total number of customers. This relation relation is known as **the Kaplan-Meier estimate of the survival function**. Let us now estimate the hazard function using the same methods.

2.3.2 Estimating the hazard function. Time-to-event data could be summarized via the hazard function. It displays the dependence of purchases on time.

Life-table estimate of the hazard function. The life-table estimate can be obtained by grouping the survival times into m intervals. Let d_j be the number of customers who made purchases in the j^{th} interval and n'_j the average number of customers in that interval. Assume that the purchase rate is constant through the j^{th} interval. The average time when no purchase is observed is $\left(n'_j - \frac{d_j}{2}\right)\tau_j$, where τ_j denotes the length of the interval. The life-table estimate of the hazard function is yielded by

$$\hat{h}(t) = \frac{d_j}{(n'_j - \frac{d_j}{2})\tau_j},$$

for $t'_j \leq t < t'_{j+1}$, $j = 1, 2, \ldots, m$.

Estimating the cumulative hazard function. The hazard function at time t, $H(t)$ has been defined from (2.2.7) to be $H(t) = -\log S(t)$. In the case of Kaplan-Meier estimate for the survival function, $S(t)$ is estimated by $\hat{S}(t)$. Subsequently, the estimate of the hazard function at time t is given by $\hat{H}(t) = -\log \hat{S}(t)$. Using the equation (2.3.3)

$$\begin{aligned} \hat{H}(t) &= -\log \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right), \\ &= -\sum_{j=1}^{k} \log \left(\frac{n_j - d_j}{n_j}\right). \end{aligned}$$

In addition, using the series expansion of $\log(1-x) = -x - x^2/2 + h.o.t$, it follows that

$$\log \left(\frac{n_j - d_j}{n_j}\right) = \log \left(1 - \frac{d_j}{n_j}\right) \approx -\frac{d_j}{n_j}$$

on neglecting terms involving second order as well as higher-order terms. Subsequently,

$$\hat{H}(t) \approx \sum_{j=1}^{k} \frac{d_j}{n_j}, \tag{2.3.4}$$

where d_j represents the number customers who purchased and n_j the total number of customers in the j^{th} interval.

This relation denotes the cumulative sum of the estimates probabilities of purchase from the first to kth interval. It is referred to as the estimate of the cumulative hazard function and is known as the **Nelson-Alain estimate of the hazard function**.

2.3.3 Comparison of two groups of survival data. In this section, two non-parametric approaches such as the log-rank and the Wilcoxon test are discussed to compare two groups of survival data.

The log-rank test. Given two groups A and B with respectively n_{1j} and n_{2j} customers at time t_j, $j = 1, 2, \ldots, r$, one is interested in the purchase times in these groups. Let r be different purchase

Table 2.1

Group	No of purchases at t_j	No of customers who didn't buy at t_j	Total
A	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
B	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

times, $t_1, t_2, \ldots, t_r$ through the two groups. Suppose d_{1j} and d_{2j} to be respectively the number of subjects from group A and B who made purchases at time t_j. The total number of purchases made at time t_j is $d_j = d_{1j} + d_{2j}$ out of $n_j = n_{1j} + n_{2j}$. This scenario is summarized in the Table 2.1

Let us consider n_{1j} to be a random variable that takes any value in the range from 0 to the minimum of d_j and n_{1j}. It turns out that d_{1j} follows a hypergeometric distribution. The probability that the random variable related to the number of purchases in the group A takes the value d_{1j} is yielded by

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}, \tag{2.3.5}$$

where

$$\binom{d_j}{d_{1j}} = \frac{d_j!}{d_{1j}!(d_j! - d_{1j}!)},$$

denotes the number of ways through which d_{1j} times could be picked from d_j. The two remaining terms in the expression (2.3.5) follow the same logic. According to the hypergeometric distribution, the mean of the random variable d_j is

$$e_{1j} = \frac{n_{1j} d_j}{n_j}. \tag{2.3.6}$$

The above expression represents the number of subjects in group A who make purchases at time t_j. This shows that the probability of making a purchase does not depend on the group the subject belongs to. It depends on the probability of purchases $\frac{d_j}{n_j}$ made at time t_j.

An overall deviation of various values of d_{1j} from the mean e_{1j} is given by

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j})$$

This expression gives the sum of differences $d_{1j} - e_{1j}$ over the number of purchase times r

Since purchase times are independent of each other, the variance of U_L is given by the sum of variances of d_{1j}. According to the hypergeometric distribution, the variance of d_{1j} is

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}. \tag{2.3.7}$$

The variance associated with U_L is

$$var(U_L) = \sum_{j=1}^{r} v_{1j} = V_L. \tag{2.3.8}$$

In addition, when the number of purchase times is large, it turns out that U_L follows approximately the normal distribution (Collett, 1994). Therefore,

$$\frac{U_L}{\sqrt{V_L}} \sim N(0, 1).$$

Typically, the square of the standard normal random variable follows a χ^2 distribution with one degree of freedom. We then have

$$\frac{U_L^2}{V_L} \sim \chi_1^2. \tag{2.3.9}$$

The above expression is known as the log-rank test.

The Wilcoxon test. It is quite similar to the log-rank test. The Wilcoxon is given by

$$U_W = \sum_{j=1}^{r} n_j (d_{1j} - e_{1j}), \tag{2.3.10}$$

where d_{1j} is defined as in the previous case and e_{1j} is given by the relation (2.3.6). The main difference is that in the Wilcoxon test, every difference is weighted by the total number of customers. This is done to give less importance to the difference $d_{1j} - e_{1j}$. It turns out that the Wilcoxon test is less sensitive than the log-rank test (Collett, 1994). The variance of U_W is given by

$$V_W = \sum_{j=1}^{r} n_j^2 v_{1j},$$

where v_{1j} is defined previously in relation (2.3.7). Finally, the Wilcoxon test follows a chi-squared distribution, we then have

$$W_W = \frac{U_W^2}{V_W} \sim \chi_1^2.$$

In general, if the hazard functions of the two groups are proportional i.e. the probability that a customer of one group makes a purchase at a given time is proportional to another one from the other group, the log-rank test is convenient to test the hypothesis that two survival functions are equal. In other cases, the Wilcoxon test statistic is performed. In practice, if two survival functions do not cross one another when plotted, this is often taken as evidence in favour of proportionality (Collett, 1994).

2.4 Proportional hazard model

When comparing two or more groups of survival data, the non-parametric methods such as log-rank and Wilcoxon test do not allow to assess the impact of several explanatory variables to the event of interest. Recall that in time-to-event data the key feature is the event of interest. The objective of modelling such data is: (i) to identify the predictor variables that are more informative about the event of interest i.e. which combination of predictors that affect the hazard function (purchase), (ii) to estimate the cumulative hazard function.

In this section, we highlight the proportional hazard (PH) model also called the Cox regression model initiated by Cox (1972) that enables to explore the relationship between different predictor variables and the event of interest.

General form of the Cox regression model. This model relies on assumption stating that the impact of predictor variables on the outcome is invariant over time. Given k predictors, $X_1, X_2, \ldots, X_k$, the general form of the Cox proportional hazard model is given by

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}), \tag{2.4.1}$$

where $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ denotes the linear component of the model. $h_0(t)$ is a non-negative function called the baseline hazard function. It is the time-dependent component of the PH model and is equivalent to the hazard function when all predictor variables for a customer are zero. $\beta_1, \beta_2, \ldots, \beta_k$ represent the coefficients of the regression model $\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$.

The model can be re-written as

$$\log \frac{h_i(t)}{h_0(t)} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \tag{2.4.2}$$

therefore, it can be referred to as a linear model for the logarithm of the hazard ratio. Notice that no assumption regarding the actual form of the baseline hazard function has been made. Although β-coefficients can be estimated independently of the baseline hazard function $h_0(t)$, we may need in some case to estimate $h_0(t)$ itself when fitting the model.

2.4.1 Fitting and interpreting the model. . To fit the model, we need to estimate two main features, the β-coefficients and the baseline hazard function $h_0(t)$. We begin to estimate the unknown β's in the linear part of the model. Once the β's are estimated they contribute to estimating the baseline hazard function. The β's are estimated via the maximum likelihood approach.

The estimates of the β-coefficients are going to measure the influence of the predictor variables on the log of the hazard ratio. As a result, $\exp(\beta)$ assesses the impact on the hazard ratio and thus on the relative probability of purchasing, where "relative probability" is interpreted as relative to the baseline hazard. The hazard for an individual with a single predictor variable is given by

$$h(t) = h_0(t) \exp(\hat{\beta} x),$$

where $\hat{\beta}$ is the estimate of the β-coefficient.

Consider hazards for two customers given respectively by

$$h_1(t) = h_0(t) \exp(\hat{\beta} x_1),$$

$$h_2(t) = h_0(t)\exp(\hat{\beta}x_2),$$

Their corresponding hazard ratio (HR) is

$$HR = \frac{h_1(t)}{h_2(t)} = \frac{h_0(t)\exp(\hat{\beta}x_1)}{h_0(t)\exp(\hat{\beta}x_2)} = \exp(\hat{\beta}(x_1 - x_2)). \tag{2.4.3}$$

When $x_1 = x_2 + 1$ the hazard ratio is simplified and takes the form $HR = \exp(\hat{\beta})$. It corresponds to the impact of an increase of 1 in the predictor variable X on the hazard. $\hat{\beta}$ is considered as a log hazard ratio since it could be expressed as $\hat{\beta} = \log(HR)$. A value of the $HR > 1$ indicates that the customer with the predictor variable x_1 has a higher probability of making a purchase than the customer with the covariate x_2. Whereas $HR < 1$ gives higher probability to the customer with x_2.

Generally, if $x_1 = x_2 + l$, $\exp(\hat{\beta}(x_1 - x_2)) = \exp(\hat{\beta}l)$, this shows that an increase of l in the predicting variable leads to change βl in the log of the HR. Subsequently, the standard error of the estimate of the log HR is then given by $l \times s.e(\hat{\beta})$. Once the model is fitted, we can check whether the PH assumption is not violated and select the best model based on the criterion discussed below.

Validation and selection of the model. The cox regression model is a semi-parametric model since it does not rely on any distributional probability assumptions. However, recall that it takes into account the proportional hazard assumption stating that the effect of different predictor variables on the event of interest remains the same over time. This is verified by (i) checking if Kaplan-Meier curves of two groups are parallel, (ii) inspecting the plot of the hazard ratio over time, (iii) running a test for the proportional hazard.

Since the model can be run under different predicting variables, we select the best model by performing a comparison between models. Two models are nested if all predictor variables of one model are included in another model. Consider

$$Mod1 : \log\frac{h(t)}{h_0(t)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_h x_h,$$

$$Mod2 : \log\frac{h(t)}{h_0(t)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_h x_h + \beta_{h+1} x_{h+1} + \cdots + \beta_k x_k,$$

$Mod1$ is said to be nested into $Mod2$. It turns out that these models are compared via the log-likelihood ratio test

$$-2\log\frac{L1}{L2} = -2(\log L1 - \log L2),$$

When two models are not nested, they are compared based on the Aikaikes Information Criterion (AIC). The best model presents the lowest AIC.

$$AIC = -2\log L + 2(k+1),$$

where k denotes the number of predictors in the model.

3. Tree-based methods

3.1 Introduction

Tree-based methods are used for regression and classification problems. The main idea is to split the predictor space into small regions. When predicting we assign the mean value or the mode from the training dataset in the region, to a new observation belonging to that region. Since the building process is summarized in a tree, these approaches are called decision tree methods. Tree-based methods are simple and easy to interpret, but they can return mediocre predictive accuracy. Hence, to enhance their predictive performance we introduce methods like bagging, random forests, and boosting. These features improve the prediction accuracy by aggregating multiples trees to produce a single aggregated prediction. However, their results can be hard to interpret.

3.2 Regression trees

The tree building process is summarized in two steps:

(i) We split P features $(X_1, X_2, \ldots, X_P)$ of the predictor space into J different regions $(R_1, R_2, \ldots, R_J)$.

(ii) Observations sharing the same region $R_j \quad (j = 1, 2, \ldots, J)$ have the same prediction, i.e. the mean value of the training dataset within R_j.

The idea is to determine J regions that minimize the residual sum of squares (RSS)

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2, \tag{3.2.1}$$

$\hat{y}_{R_j}$ is the mean value of the training data set within R_j region. Considering each partition into J regions is computationally impossible. Hence, to perform the idea above, we consider a top-down recursive binary splitting approach. Starting at the top where all observations are in the same regions, we select respectively the predictor X_j and the cut point s that minimize the RSS. At this step we create two regions $R_1(j, s) = \{X | X_j < s\}$ and $R_2(j, s) = \{X | X_j \geq s\}$ such that the value of j and s minimize the equation

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2, \tag{3.2.2}$$

where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are respectively the mean values within $R_1(j, s)$ and $R_2(j, s)$ regions. For minimizing the RSS, we repeat the same process with resulting regions till we meet a stopping criterion. Once all regions are performed, we use their mean values to predict future observations belonging to each of them.

3.2.1 Tree pruning. The approach described above is convenient for small trees. For a complex tree, it may predict well on training dataset but perform poorly on a test data set. This is due to over-fitting the data in the training dataset. To remedy this issue we develop a large tree T_0 to reduce the RSS and prune it back to get a sub-tree that minimizes the test error rate through cross-validation. Since

the tree has a great number of sub-trees, instead of taking into account each and every one, we apply a method known as cost complexity pruning (also called weakest link pruning) to select a small set of sub-trees indexed by a non-negative tuning parameter α. For each value of α, there exists a sub-tree $T \subset T_0$ such that the training error

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|, \tag{3.2.3}$$

is minimized. $|T|$ denotes the number of terminal nodes and α is the tuning parameter controlling the complexity and the fitness of the data. The more we increase α, the more the tree get pruned and hence the better is the prediction. We select the value of α that minimizes (3.2.3) using cross-validation. We then get back to the full dataset to pick the sub-tree related to α.

3.3 Classification trees

A classification tree is grown similarly to a regression tree. However, it is effective for the prediction of a qualitative response rather than a quantitative one. When predicting we allocate a given observation in the region to the most commonly occurring class of training dataset in that region. For classification trees, instead of using the RSS as a criterion of binary splitting, we are interested in the classification error rate which is the proportion of the training data set that are not from the most common class in each region.

$$E = 1 - max_k(\hat{p}_{mk}), \tag{3.3.1}$$

$\hat{p}_{mk}$ is the fraction of training dataset in the m^{th} region that are from k^{th} class. Practically, the classification error is not sufficiently sensitive for tree-growing (James et al., 2013). Hence, other splitting criteria are usually preferable, such as the Gini index and the cross-entropy respectively given by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{3.3.2}$$

and

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{3.3.3}$$

The options above are minimized for the $\hat{p}_{mk}$'s close to 0 or 1. When observations emanate predominantly from a single class, the value of these features is small. Therefore, they act as measures of node purity and can be used when pruning a tree to determine the quality of a particular split.

In general, tree-based models do not present the same level of predictive accuracy as classical supervised learning methods. However, approaches like bagging, random forests, and boosting can enhance remarkably their performance. These features are the object of the next section.

3.4 Bagging, Random forest, Boosting

These methods involve generating multiple trees resulting in a single prediction to improve the predictive performance.

3.4.1 Bagging. Predictions from a single classification or regression tree tend to have high variance. This means that if we divide the learning set randomly in half and fit separately the two proportions their resulting prediction could be different. The goal of bagging is to produce predictions with lower variance i.e. if applied to different datasets. A natural way to reduce variance is to train our method on many samples from the population and average over all predictions to produce a single low-variance model

$$\hat{f}_{av}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x). \tag{3.4.1}$$

Unfortunately, this is infeasible because we do not have multiple training datasets. To remedy this issue, we introduce the bootstrapping aggregation or simply bagging approach tailored to reduce the variance. Bagging consists of generating repeated samples from the original dataset. We then form B bootstrapped learning datasets on which we train our method and average all predictions to yield a single model

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x). \tag{3.4.2}$$

The description above is applied to regression trees. For classification trees, for each bootstrapped training dataset, we record the class predicted by any of the grown trees and consider the plurality vote i.e. the most occurring class is taken for an overall prediction.

3.4.2 Out-Of-bag Error Estimation. There exists a right way to determine the test error of a bagged tree. It has been shown that a bagged tree uses approximately two-thirds of the observations (Breiman, 1996b). The left out one-third are referred to as out-of-bag (OOB) observations. The i^{th} OOB observation is predicted using trees in which it was left out and then averaged (for a numeric response) or majority voted (for a categorical outcome) to get the response associated with a single i. The OOB prediction is computed from n OOB observations via the same process. An overall OOB MSE is derived for a quantitative response or classification error for a qualitative response. The resulting OOB error can be referred to as a reliable estimate of test error for a bagged model since it is derived from observations predicted using trees grown without them.

3.4.3 Random forests. Random forests provide a refinement of bagged trees by introducing a randomization sampling to decorrelate trees. The key idea is that an ordinary bootstrapped sample leads to the correlation of trees in a sense that if some features are strong predictors of the response they will be drawn in all bootstrapped training datasets and influence trees to be correlated. Hence, in this case averaging the result does not improve on variance. Random forests address this issue by selecting at each split a random subset of m candidates from p candidates. It has been shown (James et al., 2013) that averagely $(p-m)/p$ splits don't even consider the strong candidates. In this way, moderate predictors get the opportunity to contribute to generating decorrelated trees, resulting in predictions with lower variance. Practically, m is taken to be approximatively the square root of p ($m \approx \sqrt{p}$)

(Ishwaran et al.). It turns out that a random forest process with $m = p$ is simply a bagging process. A small value of m is typically useful if most of the predictors are correlated.

3.4.4 Boosting. The other method that improves the tree-based model predictive force is boosting. Boosting performs similarly to bagging. The only difference is that trees are grown sequentially instead of independently. Moreover, it gives misclassified observations high importance since the next model is built on the residuals of the current model. In addition, boosting usually uses small trees that allow to fit residuals and improve the learning process where it does not predict accurately and progressively updates the decision tree.

The methods introduced above are commonly used for regression (continuous response) and classification (categorical) problems. However, these methods are not immediately applicable to survival data, because they lack a mechanism for dealing for censored data. The random survival forest method that captures this feature is introduced in the following section.

3.5 Random survival forest

Random survival forests (RSF) method (Ishwaran et al., 2008) is the extension of the Breiman's random forests (RF) (Breiman, 2001). It is an ensemble tree method tailored to handle right censored as well as time-to-event issues. Random survival forests capture non-linear effects and interactions involving multiple variables. It is a totally non-parametric approach since it does not require any specific assumptions. In survival data, the time and censoring information are the key concepts. Therefore, these features are taken into account when applying the splitting rule in growing a survival tree. Deeper than that, the predictive response through the terminal nodes, the aggregated value from the forest as well as the feature assessing the prediction accuracy must involve survival information.

In this section, we begin by introducing a binary survival tree, and use its terminal nodes for predictions. The cumulative hazard function associated with each node terminal is estimated by Nelson-Alain estimator. We use the estimated cumulative hazard function from terminal nodes to determine the bootstrapped and the OOB ensemble cumulative hazard functions. We use these ensemble functions to compute the ensemble estimation of the purchase.

3.5.1 Binary survival tree. Binary survival trees are grown similarly to regression and classification trees. The process follows a binary recursive splitting through each node. A tree is grown from the root node, where all data belongs to the same node by applying the survival splitting rule. The process is the following: the root node is divided into two children nodes, the left, and the right child. In turn, the resulting children are split each one into two other children. The splitting process is followed until a stopping criterion is met, based on either node-size or node-impurity. A good split is required to maximize the difference between children nodes. We use terminal nodes for prediction.

3.5.2 Terminal node prediction. The last nodes in survival tree are called terminal nodes. Let G represent the ensemble of all terminal nodes. Assume that within the terminal node l are n customers. Suppose $(\tau_{1,l}, \delta_{1,l}), \ldots, (\tau_{n,l}, \delta_{n,l})$ are survival times and $0 - 1$ censoring information for customers associated with the terminal node $l \in G$. At a given time $\tau_{i,l}$, a customer who makes a purchase is notified by $\delta_{i,l} = 1$. A customer who does not make any purchase at that time, is said to be right censored and is notified by $\delta_{i,l} = 0$.

Suppose $t_{1,l} < t_{2,l} < \cdots < t_{N,l}$ are different purchase times. Let $d_{j,l}$ and $n_{j,l}$ be respectively the number of customers who purchased and the total number of customers associated with l node at time $t_{j,l}$.

The cumulative hazard function associated with the node l is estimated by the Nelson-Aalen estimator (2.3.4) given by

$$\hat{H}_l(t) = \sum_{t_{j,l} \leq t} \frac{d_{j,l}}{n_{j,l}}. \tag{3.5.1}$$

All observations falling into the same node l have the same cumulative hazard function.

Let $H(t|x_i)$ denotes the cumulative hazard function for a customer i with predictor features x_i. Since every observation must end up in a single terminal node $l \in G$ when applying a binary splitting rule, the cumulative hazard function for i is estimated by Nelson-Aalen estimator evaluated at the terminal node l. It follows that

$$H(t|x_i) = \hat{H}_l(t), \quad x_i \in l. \tag{3.5.2}$$

The expression above represents the cumulative hazard function for all customers as well as for the tree. However, it is associated with a single tree. RSF is a tree-based technique that aggregates several trees. Hence, we generate B bootstrapped samples from the original data and grow a tree on each sample and average their result.

3.5.3 The bootstrap and out-of-bag ensemble cumulative hazard function. Recall that Out-Of-Bag (OOB) data denotes observations non-used when growing a tree whilst in-bag (bootstrap) data are used observations. Let $J_{i,b} = 1$ denotes the case when i is left OOB when growing a tree b, and $J_{i,b} = 0$ otherwise. Let the relation (3.5.2) take the form $H_b(t|x)$ for the b^{th} grown survival tree. The OOB ensemble cumulative hazard function for i is given by

$$H_e^o(t|x_i) = \frac{\sum_{b=1}^{B} J_{i,b} H_b(t|x_i)}{\sum_{b=1}^{B} J_{i,b}}, \tag{3.5.3}$$

Since the numerator and the denominator of (3.5.3) result respectively in the sum of B bootstrapped samples over the number of bootstrapped samples, it turns out that it is an average over bootstrapped samples in the case when i is not used when growing a tree. The bootstrapped ensemble cumulative hazard function is given by

$$H_e(t|x_i) = \frac{1}{B} \sum_{b=1}^{B} H_b(t|x_i). \tag{3.5.4}$$

It turns out that the equation above is an average over bootstrapped samples taking into account both cases when i is either in-bag or out-of-bag observation. The relations (3.5.4) and (3.5.3) are used to predict outcomes. Our method has the "conservation-of-events" property (Naftel et al., 1985) that states that the sum of the estimates of the cumulative hazard functions over time is equivalent to the total number of purchases. This is valid for a range of estimators as well as for Nelson-Aalen estimator. For a given terminal node $l \in G$ in survival tree, the conservation-of-event is expressed as follows

$$\sum_{i=1}^{n(l)} \hat{H}_l(\tau_{i,l}) = \sum_{i=1}^{n(l)} \delta_{i,l} \quad \forall l \in G, \tag{3.5.5}$$

The relation above asserts that the total number of purchases is conserved within the node l. Since the number of purchase is conserved within each node, as a result it is also conserved within the survival tree.

$$\sum_{i=1}^{n} H(\tau_i|x_i) = \sum_{l\in G}\sum_{i=1}^{n(l)} \hat{H}_l(\tau_{i,l}) = \sum_{l\in G}\sum_{i=1}^{n(l)} \delta_{i,l} = \sum_{i=1}^{n} \delta_i, \tag{3.5.6}$$

where the right-hand side denotes the total number purchases within the tree.

3.5.4 Ensemble estimation of purchase. The ensemble estimation of purchase is given by the expected value of the sum of the cumulative function over time τ_k given x_i. It gives the expected number of purchases under the proportional hazard assumption stating that customers with the same x_i have the same cumulative hazard function. The purchase for i is typically given by

$$P_i = \mathbb{E}_i\left(\sum_{k=1}^{n} H(\tau_k|x_i)\right) \tag{3.5.7}$$

where $\mathbb{E}_i$ is the expected value under the proportional hazard assumption. The survival tree decision confirms the proportional hazard assumption since customers within the same terminal node have the same estimated hazard function. That estimate is taken as the ensemble estimation of purchase. For a customer i it is given by

$$\hat{P}_{e,i} = \sum_{k=1}^{n} H_e(\tau_k|x_i). \tag{3.5.8}$$

Following the same logic, the out-of-bag ensemble estimation of purchase is given by

$$\hat{P}^o_{e,i} = \sum_{j=1}^{n} H^o_e(\tau_k|x_i). \tag{3.5.9}$$

This relation is used to compute the OOB prediction error for assessing the accuracy of our model.

3.5.5 Prediction error. The predictive error is determined via the Harrell's concordance index (C-index) (Harrell et al., 1982). The C-index enables to metric the model any time and carries the censoring information (May et al., 2004). It has been adopted as a feature for evaluating the prediction performance in the survival analysis context (Kattan et al., 1998). The prediction error E_r is given by $1 - C$, where $0 \leq E_r \leq 1$. $E_r = 0.5$ denotes a prediction error no better than random sampling while $E_r = 0$ corresponds to perfect accuracy.

Steps for computing the C-index according to Ishwaran and B.Kogalur (2007):

(i) Form all possible pairs of customers in our data

(ii) Ignore pairs of customers whose shorter purchase time is censored. Ignore also pairs i and j if $T_i = T_j$ unless one of them is a purchase time i.e. either $i = 1, j = 0$ or $i = 0, j = 1$. The latter case concerns ties if one of the customers makes a purchase and the other is censored. Let $Permissible$ represent the total number of all permissible pairs

(iii) Record 1 for each permissible pair $T_i \neq T_j$ if a purchase occurs for shorter purchase time. Record 0.5 if predicted responses are tied. Let $Concordance$ be the total sum over all permissible pairs.

(iv) The C-index is given by $C = \frac{Concordance}{Permissible}$

OOB predictive error: Computing the C-index involves the predicted response (Ishwaran et al., 2008). We use the OOB estimated ensemble purchase (3.5.9) to estimate the OOB C-index C^o and subsequently the relative OOB error rate.

Recall that $\sum_{s=1}^{m} H_e^o(t_s^o|x_i)$ is the probability that a customer i makes a purchase at time t_s when i is OOB. It follows that given $t_1^o, \ldots, t_m^o$ purchase times and two customers i and k, a customer i is rated to have a higher probability of making a purchase than k if

$$\sum_{s=1}^{m} H_e^o(t_s^o|x_i) > \sum_{s=1}^{m} H_e^o(t_s^o|x_k), \tag{3.5.10}$$

The OOB C-index C^o is computed as outlined above using this rule. Finally, the OOB prediction error E_o is given by $1 - C^o$ such that $0 \leq E_o < 1$.

3.5.6 Variable importance. We inspect two methods to assess which features are strong predictors of the purchase. The variable importance (VIMP) technique relies on prediction error whilst the minimal depth approach takes into account the splitting rule.

VIMP. For a given variable, VIMP is given by the change in OOB prediction error obtained before and after permuting that variable (Breiman, 2001). Variables are selected based on their VIMP. Variables with high VIMP also have the high predictive accuracy of the outcome whilst variables with zero VIMP do not contribute anything in terms of predictive accuracy. Features with negative VIMP can improve the prediction accuracy if they are misspecified. Subsequently, the two later cases are taken as non-predictive of the outcome. Another alternative to the VIMP is the minimal depth approach.

Minimal depth. This is an order statistic feature that enables the user to ascertain the predictiveness of a variable (Ishwaran et al., 2010) by inspecting its distance to the root node. It relies on assumption that strong predictive features are those which split nodes closest to the root node where they deal with large samples. Nodes are numbered from the trunk of the tree to terminal nodes depending on their relative distance (depth) to the root node (numbered 0). The minimal depth assesses strong explanatory variables by averaging the depth of the first split for each feature within the forest. Lower values of the minimal depth indicate strong predictive variables.

3.5.7 Imputation of missing data. Two main problems met when modelling missing data are: "How does the algorithm build the model in case of missing values from the training dataset?" and "how does the algorithm manages to predict the outcome in case of missing values from test dataset?" The natural way to sort this missing data issue when dealing with linear model is either to take away the missing values or impute them before building the model. However, taking them away requires that one remove the associated columns or rows and hence leads to removing either variables or observations. This solution is quite simple but it could lead to biased result. The random forest approach of Ishwaran et al. (2008) introduces a new way of dealing with missing data, which imputes missing data adaptively as the tree is grown. At each node, before applying the splitting rule, the training dataset associated with the current node is checked for missing data. Missing data are imputed by randomly picking candidates from non-missing data associated with the node concerned. Once the missing data are sorted, the process continues as usual. This procedure is repeated until a stopping criterion is met. In the case of missing data, all values of the terminal nodes are aggregated to estimate the result of an imputed dataset. The

final imputed value for missing data will be the average of the imputed in-bag data for the continuous outcome and the most occurring imputed value for categorical outcome. This process is applied to the training dataset as well as to the test dataset.

Figure 3.1 illustrates the random survival forest process (Datema et al., 2012).

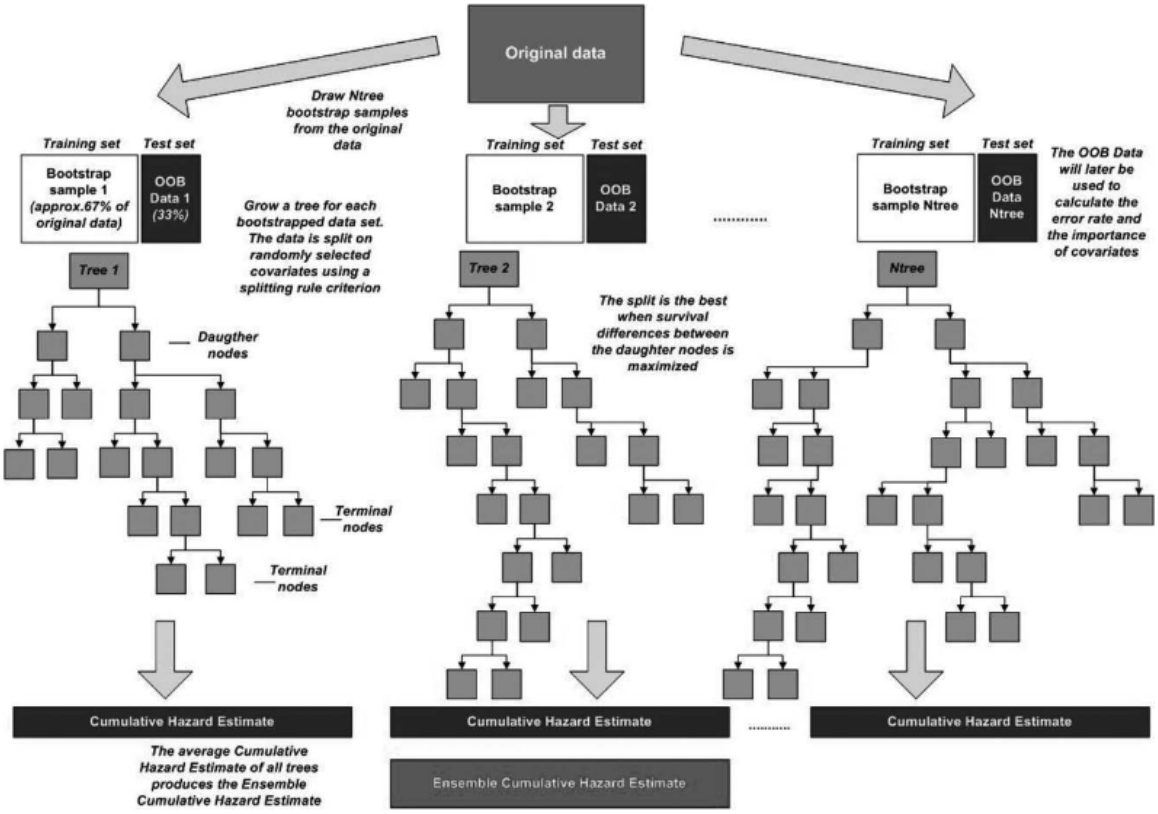


Figure 3.1: RSF process

In the next chapter, we are going to use the mathematical tools developed in two previous chapters to develop the RSF and Cox regression models. They will be used to predict probabilities of next purchase occurrence. Moreover, deeper than predicting future we also need to identify features that impact on a purchase occurrence. After that, we will assess the prediction accuracy of our models and compare outcomes from both models.

4. Application to a panel dataset

4.1 Data description and exploratory data analysis

Our data is a panel dataset collected from Australia on a sample of 2662 customers on product purchases with 35 different brands in toothpaste categories. Our data follows a sample of customers (people who have bought toothpaste in the past) for 16 weeks. For each customer, we record whether they purchase a product or brand of interest in the time period and, if so, when that purchase occurred. In addition to this data, we have the following basic information about each customer.

1) the number of purchases in the past year
2) and when the last purchase occurred
3) the average price of brand of interest for each customer
4) the average quantity of the brand
5) its average pack size
6) the customer's State
7) the number of kids in a house
8) the number of people in a house
9) the life description
10) the shopping pattern description
11) the household spend description

Figure 4.1 shows the distribution of the number of purchases in the previous year. 76.74% of consumers made purchases less ou equal to 10. The customer with the most purchases made 62 purchases and on average 7.832 purchases are assigned to each customer.

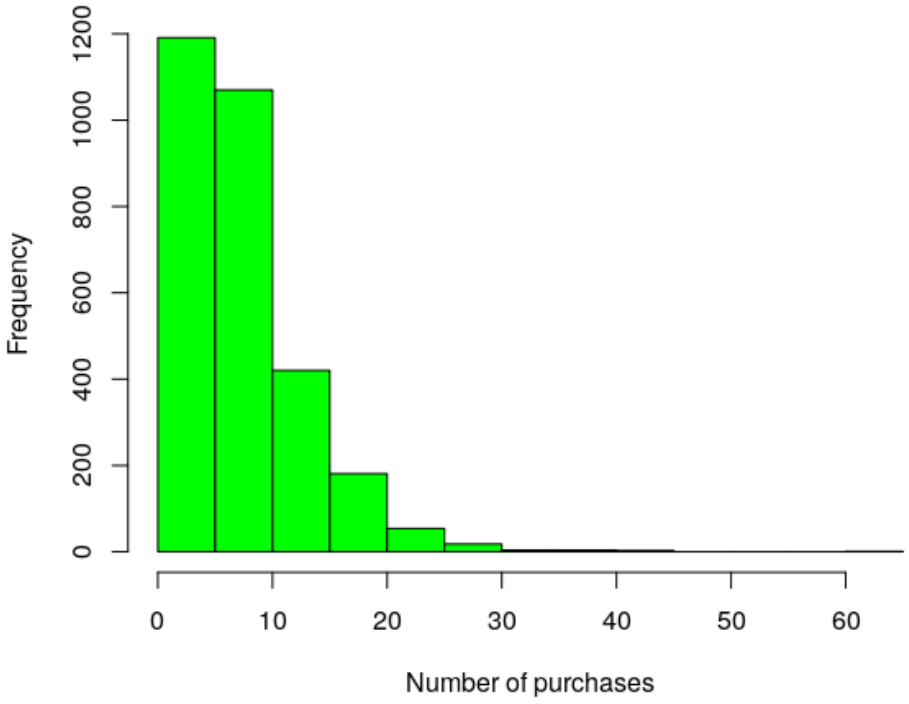


Figure 4.1: Distribution of last year purchases

4.2 Implementation details

Before fitting models, we split our dataset (2662 customers) into two subsets (different rows) called respectively **training dataset** and **test dataset**. The former is a dataset used to train our models while the latter is a dataset on which we want to make a prediction. We randomly pick approximately a half of our dataset sample (1315 customers) for training data, and the remaining (1347 customers) are considered in the test dataset. We fit both RSF and Cox regression models to the data in R software using mainly three R-packages, (i) randomSurvivalForestSRC (Version 2.4.2), (ii) survival, and (iii) Hmisc. We build models taking into account the customer's basic information aforementioned (Section 4.1), we use them as predictor variables. Based on the result, we identify factors that influence the occurrence of a purchase. They can be used when predicting risks of purchasing. For RSF, we grow the survival forest using 1000 bootstrapped samples. We select 4 predictors at each node and apply the log-rank splitting rule. We set the stopping criteria to be no fewer than 3 customers in the terminal node. We use the rank correction for censored data function (rcorr.cens) from the Hmisc package to compute the C-index used to assess the prediction accuracy of our models.

4.3 Results

In this section, we illustrate the results from both models, the predictive performance, the variable importance, and the predicted cumulative probabilities of next purchase occurrence. We give illustrations of these probabilities on some customers as well as proportions of customers.

4.3.1 Model accuracy. RSF is less accurate than Cox model. We assess both models using the C-index measure. It is used to measure the quality of the model. The value of $C = 1$ indicates a perfect prediction. $C > 0.5$ corresponds to a good prediction accuracy. $C = 0.5$ implies a predictive ability which is not better than random guessing. $C < 0.5$ implies a worse predictive ability. On training dataset, the C-index was 0.655 for RSF and 0.669 for Cox regression model. On test dataset, the C-index was 0.638 for RSF and 0.656 for Cox model.

4.3.2 Model interpretation. Both models pick the number of purchases and when the last purchase occurred as predictor variables that impact on a purchase occurrence. For training dataset, RSF shows that 1051 customers out of 1315 customers from the training sample made purchases. The error rate was less than 0.5, this shows that it is a good model. Hence, predictors used in the model are informative about a purchase occurrence.

Figure 4.2 shows the variable importance extracted from the forest. The blue colour (TRUE) indicates that the VIMP is positive while the red (FALSE) represents negative values. Recall that predictors with VIMP closer to zero or negative do not contribute anything in terms of prediction accuracy. The first two top features (the number of purchases (npch) and the last purchase occurrence (lastpch)) have larger VIMP than the remaining predictors. These predictor variables are more informative about the future purchase occurrence. We can then rely on them when predicting probabilities of observing next purchases. Intuitively, it seems obvious that the number of purchases is informative about future purchases. It shows how many times a customer attends the marketplace. Customers whose numbers of purchases are larger attend the marketplace frequently. Therefore, there is a higher probability of observing them at the market soon. Moreover, a customer should buy products depending on when the last purchase occurred. If it is a long time since that a customer has not purchased, there could be a need of purchasing again. None of the demographics were significant and the importance associated with the life description is negative.

Figure 4.3 displays the OOB error rate from the forest over the number of trees. It shows that the forest does not require a great number of trees to get the prediction error rate estimate stabilized. However, to give the opportunity to each predictor variable to contribute to the forest prediction process, a large number of random forest trees is considered.

As for the RSF, the Cox regression model also shows that 1051 customers out of 1315 customers from the training sample made purchases at least once within 16 weeks. However, we need to investigate according to this model which features are more informative when predicting next purchase occurrence. In the Cox model, the variable importance is assessed via the P-value. Based on it no significant influence on future purchase occurrence was found from the following features: average price, average quantity, average pack size, brand, state, the number of kids in the house, number of people in the house, life description, shopping pattern description, and household spent description.

The summary of the remaining significant features is given in the Table 4.1.

The null hypothesis stating that β-coefficients are zero for the number of purchases and the last purchase occurrence is strongly rejected at 0.1% significance level. According to P-value, these two features are significantly informative about purchases occurrence. For example, keeping other features constant, an

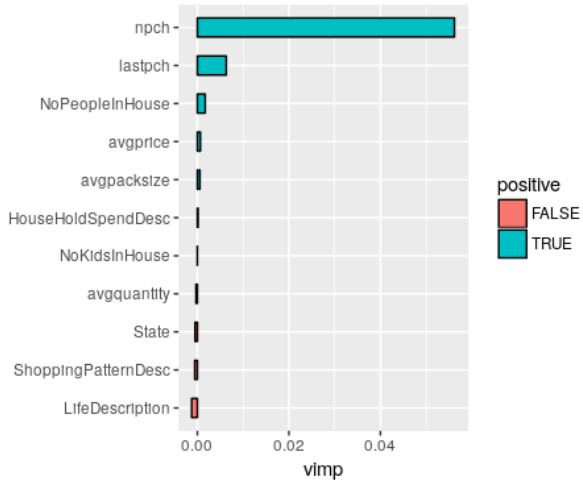


Figure 4.2: Variable importance

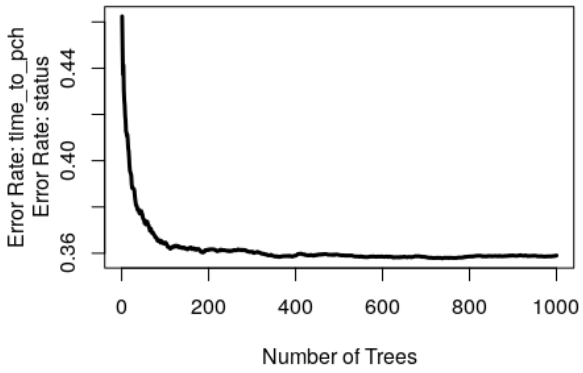


Figure 4.3: Random forest OOB generalised error

Table 4.1: Variable importance for cox model.

Covariates	Coef	exp(Coef)	se(Coef)	Z-value	P-value	95% CI
npch	0.0895010	1.0936284	0.0068125	13.138	<2e-16 ***	1.0791-1.108
lastpch	0.0096266	1.0096730	0.0029656	3.246	0.00117**	1.0038-1.016

increase of 1 in the number of purchases will increase the hazard rate of making purchases by a factor $\exp(\beta_1) = 1.09$. The third column can be referred to as the multiplicative influence of the hazard. Other tests such as Likelihood ratio test, Wald test, and Score (log- rank) were used to test the overall null hypothesis stating that the β-coefficients are not important. For each of them, the P-value was close to zero, then leading to the rejection of the null hypothesis. We conclude that variables are important when predicting next purchase occurrence.

In addition to predictive performance, the Cox model gives more information (P-value, Z-statistics, standard error (SE), the relative importance of its predictors (log ratio)) to interpret than the RSF.

4.3.3 Predicted next purchase probabilities. After training our model with the training dataset, we used it to predict the probabilities of observing next purchase occurrence using a new dataset which has not been seen by the RSF algorithm. The model shows that 1109 customers out of 1347 customers will purchase at least once within 16 weeks. The remaining are censored i.e. their purchase time goes beyond 16 weeks.

Purchase probabilities from RSF model. We extract the cumulative hazard rates $H(t)$ from the forest. The probabilities of observing next purchase occurrence are given by $h^*(t) = 1 - S(t) = 1 - \exp(-H(t))$, where $S(t) = \exp(-H(t))$ is the probability of not purchasing. It is a 1347×17 matrix, where rows represent customers and columns the follow-up time (weeks). Each column j,$(1 \leq j \leq 17)$ represents probabilities of observing purchases in week j. We can get the probability that a customer $i(1 \leq i \leq 1345)$ purchases in the week j by drawing respectively the row i and the column j. For example, the probability that a customer number 10 buys a product in the week 5 is $h^*_{10}(t_5) = 0.39$. Typically, customers don't attend the marketplace each and every week, they can sometimes take a time to use some products bought previously before buying some again. As a result, some probabilities are low others high. However, probabilities associated with customers who purchase regularly or frequently

should often be high.

Purchase probabilities from the Cox model. Once the model was trained on the training dataset, we use it to predict probabilities of the next purchase occurrence. It yields a vector of length 1345 corresponding to the number of customers. Each vector component represents the customer's hazard rate. For example, the hazard rate associated with the tenth customer is obtained by drawing the tenth component. It is $h_{10} = 0.06$. Recall that customers are followed over 16 weeks. A customer whose time purchase goes beyond 16 weeks is censored. The cumulative hazard rate associated with a customer $i(1 \leq i \leq 1345)$ in the week $j(1 \leq j \leq 17)$ is given by (2.2.6), $H_i(t_j) = h_i t_j$. For example the cumulative hazard rate related to the tenth customer in the 5^{th} is yielded by $H_{10}(t) = h_{10}t_5 = 0.06 \times 5 = 0.3$. The probability of purchasing associated with the customer i at time t_j is given by $h_i^*(t_j) = 1 - S_i(t) = 1 - \exp(-H_i(t_j)) = 1 - \exp(-h_i t_j)$, where $S_i(t_j)$ is the probability that a customer i does not buy any product up to time beyond t_j. We end up with a 1345×17 matrix, where 1345 rows denote customers and 17 (weeks) represent weekly purchase probabilities.

Illustration of the next purchase probabilities on two customers using both models. Figure 4.1 displays weekly cumulative probabilities of purchasing associated with two customers from our test sample. RSF predictions are plotted in continuous lines while predictions from the Cox regression model are plotted in dotted lines.

The first customer is plotted in blue and characterised by the following variables: number of purchases (22), last purchase (51), average price (4.41), average quantity (1), average pack size (116.9), brand (Pearl Drops), state (WA), number of kids in house (0), number of people in the house (4), life description (Older Families), shopping pattern description (Fortnightly), and household spent description ($250+). The second customer is plotted in green and has the following features: number of purchases (3), last purchase (1), average price (3.78), average quantity (1.67), average pack size (116.67), brand (Colgate), state (NSW), number of kids in house (1), number of people in the house (3), life description (Young Families), shopping pattern description (Fortnightly), and household spent description ($151-$250).

Recall that we are interested in the next purchase occurrence (the first future purchase). Both models show that the first customer i.e. plotted in blue has higher probabilities of purchasing than the second plotted in green. According to RSF, the probability that the first customer will have purchased at least once is around 0.65 after the first five weeks, around 0.68 after ten weeks and around 0.7 after fifteen weeks. For the second, it is around 0.3 after five weeks, 0.4 after ten weeks and 0.5 after fifteen weeks. For cox model, the probability that the first customer will have purchased at least once after five weeks is almost 1 and is remaining 1 all along the time interval. For the second it's around 0.65 after five weeks, 0.87 after ten weeks and around 0.95 after fifteen weeks.

This seems to confirm the piece of information given by variable importance for both models. The first customer has a great number of purchases and the last purchase compared to the second. As a result, he/she has a higher probability of purchasing sooner than the second.

Illustration of both methods using quantiles. For each model, we have cumulative next purchase probabilities at each of the 16 weeks in our study, for each customer. We summarize these weekly cumulative probabilities by plotting the quantiles of the cumulative probabilities at each week (max, upper quarter (UQ), median, lower quarter (LQ), min). Figure 4.5 displays the predicted cumulative probabilities for RSF model and Figure 4.6 concerns the Cox regression model. For RSF, the lowest and the highest probabilities in the sample were 0.005 and 0.754 whilst for the Cox regression model were respectively 0.034 and 1. The black and purple curves represent respectively the minimum and maximum probabilities observed throughout 16 weeks. The green represents the LQ (25%), the blue

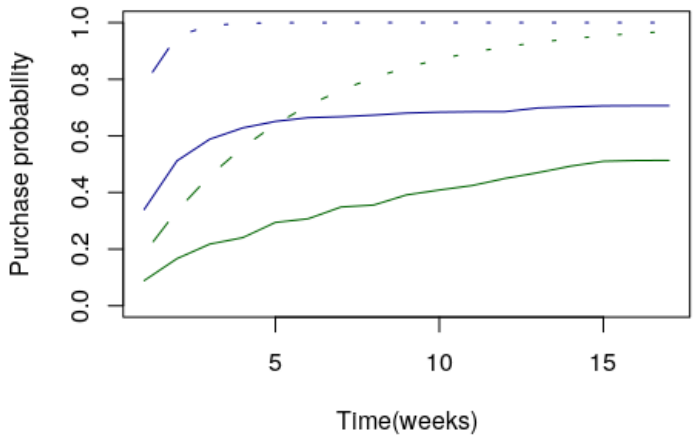


Figure 4.4: Weekly cumulative probabilities of purchasing associated with two customers (the first in blue and the second in green) predicted by the RSF (continuous line) and the Cox model (dotted line)

the median (50%), the yellow UQ (75%), and the red denotes the mean probabilities from the whole sample.

Since we are interested in the first future purchase occurrence, the green curve of Figure 4.5 illustrates that the probability that 25% of customers will have made at least one purchase is around 0.3 after five weeks, 0.5 after ten weeks and 0.6 towards fifteen weeks. The blue shows that the probability that a half of customers will have purchased at least once is around 0.4 after five weeks, 0.6 after ten weeks and 0.67 toward the end of fifteen weeks. The yellow curve is interpreted in the same way. On average, the probability that a customer purchases at least once is 0.4 at the end of 5 weeks, 0.67 at the end of ten weeks and 0.65 at the end of fifteen weeks.

Figure 4.6 describes the same scenario for the Cox regression model. As a result, curves are interpreted in the same way. It indicates that the probability that a half of the customers will have bought at least once a product is already 1 around the eighth week. Averagely, the probability that a customer purchases at least once is around 0.87 after 5 weeks, 0.94 after ten weeks, and around 0.97 at the end of fifteen weeks.

When predicting, RSF assign the same cumulative hazard function to customers sharing the same terminal node. However, some of them experience the event of interest (purchase), others do not i.e. they are censored. It gives the same cumulative probabilities of having purchased even to customers who have not purchased. As a result, the predicted cumulative probabilities cannot go up to one while for the Cox model the maximum probability is one.

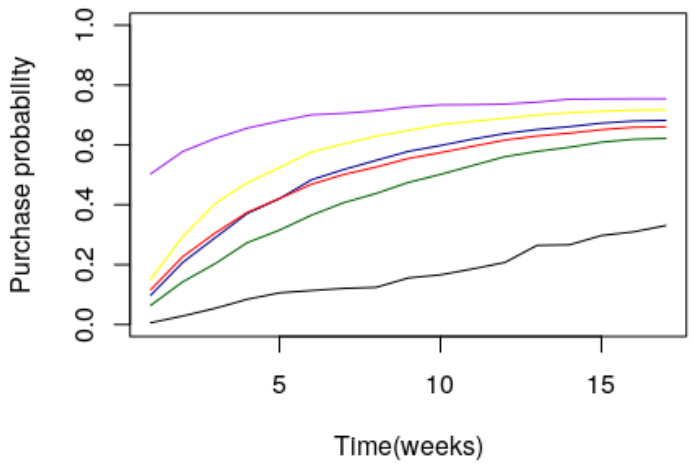


Figure 4.5: Weekly cumulative probabilities of purchasing associated with quantiles (min (black), LQ (green), median (blue),UQ (yellow), max (purple),mean (red)) predicted by the RSF model

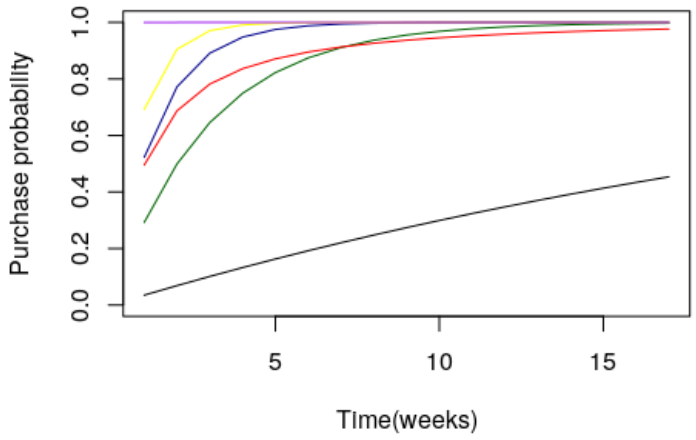


Figure 4.6: Weekly cumulative probabilities of purchasing associated with quantiles (min (black), LQ (green), median (blue),UQ (yellow), max (purple), mean (red)) predicted by the the Cox model

5. Conclusions

In the marketplace, marketers care about purchase time. Once it is known it can be used to market their products to maximize the profit. However, marketers do not know when a customer who bought a product in the past will repurchase again. In this work, we modelled probabilities of repurchasing.

We reviewed some survival analysis basics and tree-based methods. We modelled the survival data collected from Australia on purchase product in toothpaste categories. The event of interest was the purchase occurrence. People who did not buy a product by the end of the study time were censored. We applied random survival forests (Ishwaran et al., 2008) and Cox regression (Cox, 1972) models to predict probabilities that a customer who has purchased in the past will purchase again in the windows of 16 weeks.

We assessed the prediction performance using the Harrell's C-index (Harrell et al., 1982). It is a measure that people use to assess the accuracy of survival models, it roughly speaking gives the proportions of all pairs of observations in which the observation with earlier purchase time also had the higher cumulative hazard function (indicating the model got the pairwise comparison "correct", roughly speaking). According to this measure, the predictive accuracy was good for both models. Hence, they can both be used to predict probabilities of next purchase occurrence. However, the Cox regression model seems to perform slightly better than the RSF model on both training and test datasets. Moreover, it gives more information about the model, the relative importance of each predictor, the P-value as well as other statistics tests while RSF becomes like a black box in terms of the interpretability of the model.

Both models have pointed the number of purchases and when the last purchase occurred to be the more informative predictors about the future purchase occurrence. According to both methods, the higher is the probability the sooner is the purchase. A larger number of purchases implies customers who purchase frequently and hence who will attend the marketplace early. No significance influence from demographics on purchase occurrence was found.

Our models allow marketers to identify which customers are most likely to buy soon, and which customers are not likely to buy soon. They can use it by identifying (i) how frequently people shopped in the past while; (ii) when last they shopped. So they should set out to collect this information, because that tells them how they can plan for the future.

Our model gives also a possibility to a marketer to send an online remind of shopping soon to everyone with purchasing probability greater than 0.7 (for example). Or to send those with probability is less that 0.1 (just picking a number at random) a voucher giving them a discount to encourage them to purchase.

Our work was limited to predicting a single repeat purchase i.e. the first future purchase. For the future work, these methods should be extended to predict multiple repeat purchases.

Acknowledgements

My special gratitude goes to my supervisor Professor Ian Durbach for his guidance, availability, and patience throughout this research. I would like to thank the AIMS South Africa house for providing me with scientific knowledge, professional and personal guidance. I am also grateful to Prof Neil Turok for the brilliant idea of developing Africa through Sciences.

Thanks to my family especially my elder brother Ananias BUCUMI for his support, advice, and encouragement throughout my studies.

References

L. Breiman. Out-of-bag estimation, 1996b.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

D. Collett. Modelling survival data. In *Modelling Survival Data in Medical Research*, pages 53–106. Springer, 1994.

D. Collett. *Modelling survival data in medical research*. CRC press, 2015.

D. R. Cox. The analysis of multivariate binary data. *Applied statistics*, pages 113–120, 1972.

F. R. Datema, A. Moya, P. Krause, T. Bäck, L. Willmes, T. Langeveld, B. de Jong, J. Robert, and H. M. Blom. Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. *Head & neck*, 34(1):50–58, 2012.

E. Gehan. Estimating survival function from the life table. *Journal of Chronic disease*, 21:629–644, 1969.

F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

H. Ishwaran and U. B.Kogalur. Random survival forests for r. *Rnews*, 7(2):25–31, 2007.

H. Ishwaran, U. B. Kogalur, M. U. B. Kogalur, and X. Suggests. Package 'randomsurvivalforest'.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010.

H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn. Random survival forests for high-dimensional data. *Statistical analysis and data mining*, 4(1):115–132, 2011.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

M. W. Kattan, K. R. Hess, and J. R. Beck. Experiments to determine whether recursive partitioning (cart) or an artificial neural network overcomes theoretical limitations of cox proportional hazards regression. *Computers and biomedical research*, 31(5):363–373, 1998.

M. May, P. Royston, M. Egger, A. C. Justice, and J. A. Sterne. Development and validation of a prognostic model for survival time data: application to prognosis of hiv positive patients treated with antiretroviral therapy. *Statistics in medicine*, 23(15):2375–2398, 2004.

D. Naftel, E. Blackstone, and M. Turner. Conservation of events. *Unpublished notes*, 1985.

R. Peto, M. Pike, P. Armitage, N. E. Breslow, D. Cox, S. Howard, N. Mantel, K. McPherson, J. Peto, and P. Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British journal of cancer*, 35(1):1, 1977.