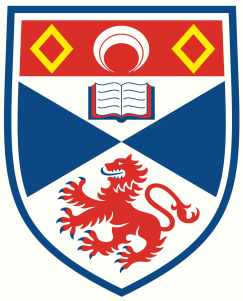
Modelling the prognosis from colorectal cancer on the basis of image analyses

Man Ho Suen

Supervised by Dr Peter David Caie and Dr Giorgos Minas



University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Master of Science in Statistics

at the University of St Andrews

August 2019

Abstract

This dissertation presents an analysis of colorectal cancer (CRC) data with clinico-pathological and spatial density distribution of macrophage, lymphocytic infiltrates and tumour buds (TBs) variables. This study aims to build a survival model and a classification rule, in order to enhance diagnosis and prediction of CRC survival. The data contain three cohorts, two from Edinburgh and one from Japan. The one from the Edinburgh served as the training dataset while the rest allowed completely unseen validation on the fitted models.

For survival analysis, Cox's models incorporating regularisation and machine learning techniques were used in view of the high dimensionality and collinearity of the data. Due to the class imbalance of the data, oversampling was implemented to compare if any improvements on the prediction performance. Generalised Boosted Regression Models (gbm) performed the best with Harrell's Concordance index (C-index) of approximately 0.7 for both validation cohorts.

For classification analysis, the survival responses were reduced to a binary outcome, high and low risk group. In view of the non-linearity, machine learning techniques, namely Naïve Bayes (NB), Support-Vector Machines (SVM) and gbm models were fitted. Their performances were compared to the performance of pT staging, which is one of the standard cancer classifications. The gbm model again achieved the best predictability, in terms of the log-rank test and Kaplan-Meier (KM) plots. NB with only one predictor related to the interaction of macrophage and lymphocytic infiltrates performed surprisingly well in the unseen Edinburgh validation cohort. This suggests the promising potential of investigating their interaction further.

Keywords: colorectal cancer (CRC), image analysis, regularisation, ridge, lasso, elastic net, adaptive lasso, machine learning, Random Forest (RF), Generalised Boosted Regression Models (gbm), Naïve Bayes (NB), Support-Vector Machines (SVM), Kaplan-Meier (KM), log-rank test.

Declaration

I, Man Ho Suen, hereby certify that this dissertation, which is approximately 14,000 words in length, has been composed by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a degree.

This project was conducted by me at the University of St Andrews from May 2019 to August 2019 towards fulfilment of the requirements of the University of St Andrews for the degree of MSc Statistics under the supervision of Dr Peter David Caie and Dr Giorgos Minas.

16 August 2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

AUROC Area Under the Receiver Operating Characteristics

C-index Harrell's Concordance index

CI Confidence Interval

CRC colorectal cancer

CT Core Tumour

CV Cross Validation

DSD Disease Specific Death

DSS Disease Specific Survival

edf effective degrees of freedom

EMLVI extramural lymphovascular invasion

FDA Flexible Discriminant Analysis

gbm Generalized Boosted Regression Models

HR Hazard Ratio

i.i.d. independent and identically distributed

IM Invasive Margin

k-NN k-nearest neighbor

KM Kaplan-Meier

LAR Least Angle Regression

lasso least absolute shrinkage and selection operator

NB Naïve Bayes

NDMCH National Defence Medical College Hospital

NSPP Neyman-Scott point process

OLS Ordinary Least Squares

OOB out-of-bag

PLDA Penalised Linear Discriminant Analysis

PRSS penalized residual sum of squares

QDA quadratic discriminant analysis

RF Random Forest

se standard error

SMOTE Synthetic Minority Over-sampling Technique

SVM Support-Vector Machines

TBs tumour buds

TME tumour micro-environment

TNM tumour-node-metastasis

WTS whole tumour section

CHAPTER ONE

INTRODUCTION

1.1 Background

Colorectal cancer (CRC) was the fourth most common cancer and the second most common cause of cancer death in the United Kingdom in 2016; there were more than 42,000 new cases and an average of around 16,000 deaths related to colorectal cancer (CRC), from 2014 to 2016 [1]. Among CRC patients, as the risk factors vary, the mortality rate is hard to predict [1]. This study will use statistical modelling to help us assess more accurately the prognosis of CRC patients.

The current practice for cancer prognosis is the widely recognised tumour-node-metastasis (TNM) staging, which categorises how the cancer spreads inside a body. Patient with stage 1 CRC cancer will have over 95% survival rate for five years or more after diagnosis; stage 4 patients have less than 10% [1]. Although TNM staging remains to be the gold standard, there is a gap in our knowledge to identify potential stage 2 patients with high risk. A further sub-categorisation of stage 2 patients is crucial to better determine suitable treatment. This opportune treatment can avoid unnecessary patient suffering.

In an attempt to fill this gap, this study analyses data based on a retrospective cohort study that mainly involves stage 1 and 2 CRC patients under the TNM standard. This cohort was chosen not only because some of these patients are at high risk if appropriate and timely treatment cannot be delivered, but also because it is more cost-effective to focus on these patients, in terms of enhancing patients' survival.

In addition to the widely recognised TNM staging, Immunoscore is one of the recent prognostic tools used for CRC [2] to avoid adjuvant treatment for stage 2 patients. This score brings the host immune system response into cancer classification. High Immunoscore suggests good prognostic CRC patients and vice versa. However, it conducts a density analysis only on lymphocytic

infiltrates (part of the host immune system) without taking the interaction with tumour buds (TBs) (individual or a small cluster of tumour cells) into account.

With advances in the technology of medical imaging, more in-depth spatial image analysis has become available for cancer prognosis. Image analysis on TBs plays a pivotal role in offering an objective perspective for Disease Specific Survival (DSS) since these samples within the patient's tissue section are hypothesised to represent the aggressive sub-type of the cancer cells. The process for manual visual inspection of TBs is tedious and subjective. Automated image analysis can be informative for contributing a more objective view on patient cancer staging.

This study will concentrate on the spatial interactions among TBs, lymphocytic and macrophage infiltrations. A TB is defined as one or a cluster of up to four tumour cells. TBs are regarded as illustrative cancer samples and directly linked to metastasis. Lymphocytic and macrophage infiltrates represent the specific and non-specific immune responses respectively. In this regard, lymphocytic infiltrates $CD3^{+}$ and $CD8^{+}$ and macrophage infiltrates $CD68^{+}CD163^{-}$ (M1 phenotype) and $CD68^{+}CD163^{+}$ (M2 phenotype) are the main focus in the data. Immunofluorescence staining allows us to visualise the spatial distribution of multiple infiltrates and TBs in the same preparation. Hence, this is unlike the traditional methods that have to work on different slides of the tissue section. These new methods provide better quality of the spatial data. The microscopic distribution mirrors how the host immune system reacts towards the cancer cells. Spatial statistics, quantified in terms of distance, quantity, density and intensity at distinct margins of the whole tumour section (WTS), can be extracted. Through statistical modelling, this data can disclose how specific types of immune infiltrates are linked to cancer prognosis.

In addition, research has been done on the spatial interaction among multiple types of immune infiltrates and TBs on tissue sections. This spatial information can be combined with clinicopathological characterisation, and can in turn improve the accuracy of the survival prediction and prospective treatment selection. This combination is likely to prove valuable in any newly developed prognostic tool.

This study will conduct survival and classification statistical analysis. It is hoped that the results can be generalised to other epithelial cancer, such as bladder cancers, for the next step. Consequently, treatment selection can be made on statistical grounds and medical resources can be allocated more efficiently.

1.2 Previous Works

This study builds on previous attempts to propose statistical models incorporating image analysis data. Nearchou *et al.* uses a combination of a Cox proportional hazard regression penalised by the lasso method (see section 2.2.2) and the Random Forest (RF) decision tree model [3].

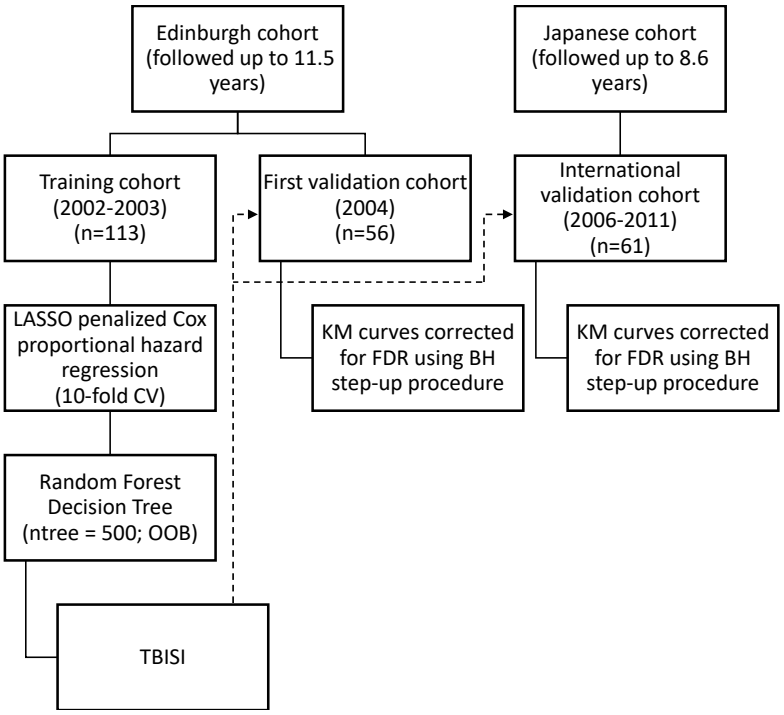


Figure 1.1: Nearchou *et al.* Statistical Analysis Workflow.

They focused on the density of certain lymphocytic infiltrates, i.e. $CD3^+$ and $CD8^+$ T cells in the vicinity of the TBs. This piece of work could be seen as a stepping stone to other immune infiltrates. Nearchou *et al.* studied the spatial interaction between those infiltrates and TBs. The proposed lasso penalised Cox model attempted to obtain a parsimonious and interpretable model by shrinking some of the predictors towards zero [4]. It was followed by a RF decision tree model that further narrowed down these selected predictors to four, by ranking the mean decrease Gini coefficient (a measure of variable importance). RF theoretically does not overfit the model thanks to the bagging algorithm (see section 2.3.1) [5]. Despite using a simple model, the RF was shown to be superior to the Immunoscore in the validation [3].

1.3 Challenges

There are several challenges in analysing the data.

First, the data is high dimensional and collinear. In particular, the spatial correlations between the data that describe the density of lymphocytic and macrophage infiltrates are very prominent. It is known that lasso Cox model gives inconsistent feature selections under these circumstances [6–8].

Second, there are substantial non-linear, complicated interactions among immune infiltrates and TBs that can not be ignored.

Third, as most of the patients are censored, the class imbalance of this data should be accommodated. There may not be sufficient data for the patients who died, which is common in medical data. This poses a problem for survival prediction and classification since the model built on imbalanced data is likely to be biased to the majority class.

1.4 Objectives

This thesis aims to model the prognosis for CRC on the basis of image analysis.

First, the high dimensionality and collinearity of the predictors have to be addressed. It is crucial to identify the influential predictors in order to be cost-effective and efficient in extracting this data in the future. It is also important to assess how well the predictors can explain the response after the unavoidable information loss due to feature selection. This study seeks statistical modelling methods to stabilise the inconsistency in lasso Cox's model observed in Nearchou *et al.*'s paper while extracting the influential predictors that explain the heterogeneity of the spatial distribution [3]. The spatial predictors in this study describe the immune infiltrates within the tumour micro-environment (TME). Statistical methods will be used in overcoming the imbalance of data.

Second, survival analysis is implemented to provide prognosis on CRC patients. This analyses the relationship between the survival responses, i.e. time-and-event response, and those highly correlated predictors. This targets to give a survival model that predicts the risk for CRC patients.

Third, classification analysis is undertaken to find classification rules that can distinguish patients into high and low risk group. This would allow us to find the predictors or combinations of predictors that characterise the differences between the two groups.

Last, as the dataset used in this study is the same as the one used in Nearchou *et al.*'s paper but with extra variables linked to macrophage infiltrates [3], this study should provide novel insights into their interactions with lymphocytic infiltrates and TBs; which has not been much discussed in recent literature.

1.5 Organisation of Dissertation

A short description of each chapter in this thesis is listed below.

Chapter 1 gives an introduction.

Chapter 2 presents the literature review on the methods applied for the analysis.

Chapter 3 describes and discusses the analysis and results.

Chapter 4 concludes the findings and limitations.

Appendix A covers the R script, the console outputs and the session information.

Appendix B provides the supplementary outputs for readers' interest.

CHAPTER TWO

METHODS

This chapter discusses the statistical models and methods used.

2.1 Ordinary Least Square and Best Subset Selection

Consider a simple linear model setting,

$$\mathbf{y} = \sum_{j=1}^{p} \mathbf{x_j}\beta_j + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ is the vector of the responses of n subjects, $\mathbf{x_j} = (x_{1j}, ..., x_{nj})^T$ for $j = 1, ..., p$ is the vector of the linear predictors, β_j is the j^{th} coefficient, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n) \sim \mathcal{N}(0, \sigma_\varepsilon)$ is the vector of the error terms. The predictors are first standardised to have zero mean and standard deviation of one before putting into the model. Standardisation transforms predictors into comparable scales otherwise predictors with larger ranges will have greater influences during modelling. Considering that not all the variables in this study are in the same scale, it is sensible to standardise them.

The vector of coefficients can be estimated, i.e. $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, ..., \hat{\beta}_p]$, by minimising the Ordinary Least Squares (OLS) equation. The OLS estimates of $\boldsymbol{\beta}$ are,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ols} &= \underset{\beta}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 \right\} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \end{aligned} \tag{2.2}$$

where $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_p}]$ and $\hat{\boldsymbol{\beta}}_{ols}$ is a vector of OLS estimated coefficients (see Proof B.2, Appendix B). However, the predictability and the interpretability in OLS are often criticised for overfitting

the train data, which is further discussed in the next paragraph [5]. The expected prediction error, also known as the generalisation error, is given by

$$\mathbb{E}\left\{(\mathbf{y}-\hat{\mathbf{y}})^2\right\} = \sigma_\varepsilon^2 + (Bias\{\hat{\mathbf{y}}\})^2 + Var\{\hat{\mathbf{y}}\}, \tag{2.3}$$

where $\hat{\mathbf{y}} = \hat{f}(\mathbf{x}^{'}) = \sum_{j=1}^{p} \mathbf{x}_j^{'}\hat{\beta}_j$ is the vector of the fitted responses and $x_j^{'}$ is the j$^{\text{th}}$ linear predictor in the new data. For derivation of equation 2.3 (see proof B.1, Appendix B).

OLS estimators are theoretically unbiased if

(i) selected subset is the underlying linear model,

(ii) homoskedasticity (share common σ_ε^2) and,

(iii) independent and identically distributed (i.i.d.) errors ($Covariance(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$);

which are hardly met in reality, in particular (iii) for spatial data. Subsequently, OLS tends to include more predictors than the underlying , i.e. overfits the training data, and hence does not generalise. Consequently, OLS often results in low bias, high variance (see equation 2.3). Moreover, collinearity causes $\boldsymbol{X}^T\boldsymbol{X}$ to be close to singular, and thus the variance of the estimator is large (see Proof B.3, Appendix B),

$$Var(\hat{\beta}_j) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma_\varepsilon^2. \tag{2.4}$$

Hence, even large coefficients may not reject the null hypothesis, i.e $\hat{\beta}_j = 0$, in significance testing.

2.2 Regularisation

In view of the high dimensional and highly correlated data in this study, best subset selection using the unbiased OLS method is computer-intensive; while any step-wise selection is only step-wise optimal. Nonetheless, the best subset selection problem is to solve the non-convex problem:

$$\hat{\boldsymbol{\beta}}_{best\ subset} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 \right\} \text{ subject to } \|\beta\|_0 \leq k, \tag{2.5}$$

where $\|\beta\|_0 = \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\}$ is the ℓ_0 norm; $\mathbb{1}\{\beta_j \neq 0\} = \begin{cases} 0, & \text{otherwise} \\ 1, & \text{if } \beta_j \neq 0 \end{cases}$ is an indicator function. Thus it is a discrete process that at most k variables are kept and $p-k$ ones are removed. It can be interpreted as a ℓ_0 penalty term that imposes explicit constraints on the

variables. However, there are more potential predictors than the number of observations ($p > n$) in the training dataset. Hence, $\mathbf{X}^T\mathbf{X}$ in equation 2.2 becomes singular [5].

Hence, shrinkage is applied to introduce bias into statistical modelling in the hope that this is compensated by a lower variance, in order to achieve more accurate predictions and more interpretable variable selection. This bias is incorporated through various forms of penalty terms. Penalised regression can be expressed in a general form,

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 + \lambda \sum_{j=1}^{p} P(\beta_j) \right\}, \tag{2.6}$$

where λ is a shrinkage parameter and $P(\beta_j)$ is a penalty function on β_j.

The shrinkage parameter, also known as the penalty parameter λ is determined by ten-fold Cross Validation (CV) in this study. CV is an objective measure to choose the best value of parameter λ since this λ best performs in the unseen data and thus is expected to generalise for prediction [7]. Ten-fold CV randomly splits the training data into ten mutually exclusive sets of approximately equal portion. It then uses nine of those portions as modelling set and the remaining portion as validation set. This cycle is iterated until all sets have been used as validation and omitted from corresponding modelling once. A general form of ten-fold cross validation formula is

$$CV(\hat{f}, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{L}(y_i, \hat{f}^{-k}(x_i, \lambda)), \tag{2.7}$$

where $\mathbf{L}$ is a loss function from a fitted model trained without the k^{th} partition of data ($\hat{f}^{-k}$) and λ is a penalty term to be tuned [7]. In each iteration, this loss function ($\mathbf{L}$) is evaluated and at the end is averaged to generate the CV error ($CV(\hat{f}, \lambda)$. The optimal $\hat{\lambda}$ that minimizes the CV error ($CV(\hat{f}, \lambda)$), is obtained through this process.

The shrinkage methods covered in this study are ridge, lasso, elastic net and adaptive lasso.

2.2.1 Ridge

The motivation of ridge regression is to coerce $\mathbf{X}^T\mathbf{X}$ to invertible [9]. Ridge regression minimises the penalized residual sum of squares (PRSS) through

$$\hat{\boldsymbol{\beta}}_{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2.8}$$

where $\lambda > 0$ [9]. Equation 2.8 solves a convex problem with ℓ_2 norm (β_j^2).

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I_{pp}})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.9}$$

$$= ((1+\lambda)^{-1})\hat{\boldsymbol{\beta}}_{ols} \tag{2.10}$$

Equation 2.9 shows that ridge regression brings in bias by imposing a penalty term with parameter λ. In the cases of orthonormal inputs, equation 2.10 illustrates how ridge estimators relate to OLS. A proportional shrinkage is applied to the OLS estimators with a denominator of $(1+\lambda)$. In other words, shrinkage is proportional to the sizes of coefficients. Owing to proportional shrinkage, standardisation is required before undergoing ridge regression; unless coefficients with larger scales are favoured for justified reasons. Furthermore, from equation 2.10, ridge regression can only shrink estimated coefficients to almost zero, i.e. cannot perform variable selection. In other words, ridge regression aims for a dense solution, i.e. most of the estimated coefficients are non-zero. For derivation of equation 2.10, see proof B.4, Appendix B.

Ridge regression inhibits large coefficients by putting up a ℓ_2 norm constraint. In contrast to OLS, all the correlated variables are penalised through proportional shrinkage because these redundant variables can be deflated without affecting the squared errors ($\|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_\mathbf{j}\|^2$). The parameter λ controls the size of penalty; the heavier the penalty, the larger the shrinkage is.

For OLS, the degree of freedom is straightforward that is equal to the number of selected variables. For regularised regressions, the effective degrees of freedom is defined as

$$\begin{aligned} df(\lambda) &= tr[\boldsymbol{X}(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I_{pp}}^{-1}\boldsymbol{X}^T] \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}, \end{aligned} \tag{2.11}$$

where d_j are the singular values of $\boldsymbol{X}$ (see proof B.5) [5]. This can be interpreted as non-zero coefficients in ridge are fitted in a regularised manner controlled by λ, i.e. effective degrees of freedom (edf) tends to 0 as λ goes to infinity.

From the Bayesian perspective, ridge regression has a multivariate Gaussian i.i.d. prior for coefficients β and inverse gamma distribution with sharp parameter α_0 and scale parameter β_0 for σ^2

$$\beta|\sigma^2 \sim N(\mathbf{0}_p, \frac{\sigma^2}{\lambda}\mathbf{I}_{pp}) \; and \; \sigma^2 \sim Inv\ Gamma(\alpha_0, \beta_0) \tag{2.12}$$

In equation 2.12, β is a vector of random variables and λ determines how certain the prior is. In other words, the larger the penalty, the more informative the prior is. Credible intervals for β can be obtained during the process, while confidence intervals for β are generated through

bootstrapping in the frequentist approach.

Despite that ridge regression forces coefficients to take values close to zero, the coefficients are not removed from the model and therefore a further analysis is necessary for variable selection.

2.2.2 Lasso

The motivation of Lasso is to achieve a sparse solution for the underlying model with ℓ_1 penalty. In contrast to ridge regression, which is based on the assumption of a dense solution, lasso presumes that there are few non-zero estimated coefficients in a high dimensional problem. In this regard, it is important to understand the "Bet on Sparsity" principle [5]. In essence, Tibshirani *et al.* suggests that sparse solution often outperforms dense one, due to the curse of dimensionality [5]. The data size required to support statistical significance grows exponentially with the dimensionality. Nevertheless, in most of the time, there are not enough data to solve a dense problem ($p > n$). Hence, sparse solution should be the goal for dimensionality reduction.

Lasso penalises through ℓ_1 norm ($|\beta_j|$) and is defined as

$$\hat{\boldsymbol{\beta}}_{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.13}$$

where $\lambda > 0$ [10]. This is also known as basis pursuit in signal processing [11].

The major difference between ridge and lasso regression is that lasso can shrink estimated coefficient ($\hat{\beta}$) to zero, while ridge regression cannot. In other words, lasso can undertake both shrinkage and automatic variable selection simultaneously. This variance stems from the distinctive natures of ℓ_1 and ℓ_2 norms. Through solving equation 2.13 in the cases of orthonormal inputs, lasso estimators can be expressed in terms of OLS [5]:

$$\hat{\boldsymbol{\beta}}_{lasso} = sign(\hat{\boldsymbol{\beta}}_{ols})(|\hat{\boldsymbol{\beta}}_{ols}| - \lambda)_+ \tag{2.14}$$

Comparing equations 2.10 and 2.14, as λ grows, both estimators shrink in different manners. Lasso shrinks variables irrespective of their sizes of coefficients, i.e. almost equal shrinkage. For derivation of equation 2.14, see proof B.8, Appendix B. The ridge ones can deflate close to zero while the lasso ones can drop to exactly zero, i.e. feature selection. Hence, lasso looks for sparse solutions, i.e. some of the estimators β_j should be zero. This enhances the interpretability of the final model though sparsity, which is unlikely in reality.

Lasso corresponds to the Bayesian estimators ($\beta|\sigma^2$) with a i.i.d. Laplace distribution, also known as double exponential distribution, prior; where σ^2 is a non-informative and scale-invariant

prior [12].

$$\beta|\sigma^2 \sim Laplace(\mathbf{0}_p, \frac{\sqrt{\sigma^2}}{\lambda}\mathbf{I}_{pp}) \; and \; \sigma^2 \propto \frac{1}{\sigma^2} \tag{2.15}$$

The structures of equations 2.15 and 2.12 are alike that the vector of β has a mean of zero and λ controls how informative the prior is. The difference is that Laplace distribution is more spiky at zero and has heavier tail than Gaussian. Consequently, lasso tends to produce either zero or large coefficient estimates.

Credible intervals for β can be obtained easily under the Bayesian framework, while the Confidence Interval (CI) for β are formulated through bootstrapping in the frequentist approach. Both the credible intervals and CI are biased and unstable. The reason for the instability is that the estimated coefficients in classical lasso are heavily dependent on how the train data is randomly split during CV, and fluctuate accordingly. In the case of correlated variables, the coefficients for these bootstrap samples can differ remarkably across bootstrap samples. The confidence intervals obtained through bootstrapping are unreliable and can be difficult to interpret since different sets of estimated coefficients can be set towards zero in lasso, with respect to the corresponding bootstrap sample.

Contrary to ridge regression, lasso aims at achieving parsimony despite the optimistic assumption about sparse solutions. In reality, correlated predictors are often encountered. In respect to these predictors, lasso selects one and eliminates the rest, i.e. fails to implement grouped selection. It can be explained by considering the trade-off in equation 2.13. It is sensible to remove these predictors since they do not contribute much in minimizing the PRSS. However, this is likely to undermine the prediction performance. In contrast, ridge regression undergoes proportional shrinkage which leads to grouping effect. This effect can be shown in equation 2.8) that highly collinear predictors have similar estimated coefficients (with a change of sign if negatively correlated). Therefore, ridge regression tends to outperform lasso in terms of predictability for correlated predictors.

Moreover, studies have shown that lasso does not always achieve oracle properties [8, 13, 14]. An oracle procedure is defined as identifying the right subset of underlying variables and having a optimal estimation rate [13]. First, criticisms were drawn on the inconsistent feature selection of lasso at optimal estimate rate, i.e. optimal λ [14]. In the particular cases of collinear variables, lasso feature selection is comparatively variable since this ability disables lasso from choosing a group of correlated variables into the model, even the optimal λ is chosen through CV. Consequently, erratic results are likely to happen since different splits of CV generate distinctive sets of predictors among those correlated groups. Second, even the right subset is chosen, lasso may have suboptimal prediction owing to biased large coefficient estimates [13].

To study the robustness of lasso variables, CI through non-parametric bootstrapping or credible intervals through Bayesian lasso can serve for post-inference. There are other ways to estimate p-value and CI with Least Angle Regression (LAR) for lasso with fixed λ [7]. The package `selectiveInference` in `R` implements the parametric bootstrapping algorithm [15, 16]

Another concern is that lasso can only select at most n variables when the number of parameters is larger than that of the observations ($p > n$). Given that it is a convex optimization problem and the rank of the matrix $\mathbf{X}$ is n, there are at most n solutions, i.e. estimated coefficients.

2.2.3 Generalisation of Lasso

Simulation studies show that lasso empirically does not cope well with high dimensionality and collinearity [6–8]. It is ascribed to the nature of ℓ_1 norm, as mentioned above. In view of these limitations, lasso coefficients may not truly reflect the relative importance of the variables. Thus, there are generalisations of lasso to improve consistency and stability. Elastic net imposes ridge penalty to alleviate collinearity while adaptive lasso tries to adjust the degree of bias.

2.2.3.1 Elastic net

Elastic net estimates $\hat{\boldsymbol{\beta}}_{enet}$ are given by

$$\hat{\boldsymbol{\beta}}_{enet} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 + \lambda \sum_{j=1}^{p} \left[\alpha|\beta_j| + \frac{1}{2}(1-\alpha)\beta_j^2 \right] \right\}, \tag{2.16}$$

where elastic net parameter $\alpha \in [0, 1]$ and $\lambda > 0$ [6]. When $\alpha = 0$, it reduces to ridge regression; when $\alpha = 1$, it is simplified to lasso regression.

The motivation for elastic net is to improve error over lasso. It strikes a balance between predictability and interpretability by adopting an extra parameter, α. It allows a continuum of models between ridge and lasso by manipulating a flexible penalty via a convex combination of ℓ_1 and ℓ_2 norm. As a result, elastic net retains the properties of ridge and lasso regressions. ℓ_1 norm enables features selection while ℓ_2 norm gives the grouping effect. By grouping correlated covariates together, this allows elastic net to perform better over lasso in terms of predictability.

The grouping effect of elastic net with identical predictors, as an extreme case, can be proved by showing equation 2.16 strictly convex for $\alpha < 1$ and $\lambda > 0$, i.e. not lasso [6]. Conversely, lasso problem is not strictly convex and may not have a unique solution when $p > n$ [17]. Therefore, another potential feature is that elastic net can have more than n non-zero coefficients when $p > n$.

In the cases of orthonormal inputs, elastic net can be expressed in terms of OLS:

$$\hat{\boldsymbol{\beta}}_{enet} = sign(\hat{\boldsymbol{\beta}}_{ols}) \frac{(|\hat{\boldsymbol{\beta}}_{ols}| - \frac{\lambda_1}{2})_+}{1 + \frac{\lambda_2}{2}}, \tag{2.17}$$

where $\lambda_1 = \alpha\lambda$ and $\lambda_2 = (1-\alpha)\lambda$ [6]. By recalling the ridge and lasso, equation 2.17 can be decomposed into two parts : $(1+\frac{\lambda_2}{2})^{-1}\hat{\boldsymbol{\beta}}_{ols}$ and $sign(\hat{\boldsymbol{\beta}}_{ols})(|\hat{\boldsymbol{\beta}}_{ols}| - \frac{\lambda_1}{2})_+$ correspond to ridge and lasso respectively.

Under the Bayesian framework, elastic net has the following priors:

$$\beta|\sigma^2 \sim \exp\left\{(-\frac{1}{2\sigma^2}(\lambda_1|\beta| + \lambda_2\beta^2)\right\} \textit{ and } \sigma^2 \propto \frac{1}{\sigma^2}, \tag{2.18}$$

where λ_1 and λ_2 are shrinkage parameters for ℓ_1 and ℓ_2 norm respectively [6]. λ_1 and λ_2 controls the proportion of lasso and ridge regressions respectively. This can be viewed as a hybrid of ridge and lasso priors, where α manipulates the spikiness and the heaviness of tail for the prior distribution.

The parameter α is often pre-assigned. Empirically, elastic net achieves satisfactory performance unless α is adjusted close to either ridge or lasso regression. This is partly due to double shrinkage that elastic net procedure can be broken down into two stages [6]:

1. Finding the ridge coefficient estimates for each fixed λ_2;

2. Implementing lasso shrinkage along the coefficient paths.

This usually does not reduce variance and entails unnecessary bias. To alleviate this problem, Zou suggested rescaling the parameter estimates by adding variance with a denominator $(1+\lambda_2)$ [6]. Nevertheless, it is not implemented in the `glmnet` package in `R` [18].

2.2.3.2 Adaptive Lasso

Adaptive lasso is aspired to adjust lasso, so as to fulfil the oracle properties. The adaptive lasso estimates $\hat{\boldsymbol{\beta}}_{alasso}$ are defined as

$$\hat{\boldsymbol{\beta}}_{alasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 + \lambda_n \sum_{j=1}^{p} w_j|\beta_j| \right\}, \tag{2.19}$$

where $w_j = \frac{1}{|\tilde{\beta}_j|^\gamma}$ is a weight for a pilot estimate $\tilde{\beta}_j$, $\gamma > 0$ is a constant that adjusts the weight (w_j) and λ_n varies with n [8]. Typically, possible values for γ are 0.5, 1 and 2. The pilot estimate

($\tilde{\beta}_j$) can be obtained from OLS for $p < n$ [8], and univariate OLS for $p > n$ [19]; ridge estimates are more often applied for simplicity [8]. In the cases of $\tilde{\beta}_j = 0$, the weights go to infinity, i.e. $w_j = \infty$, and thus the estimated coefficient $\hat{\beta}_j = 0$. This weighting scheme can be extended to ridge and elastic net [20]. It could be seen from equation 2.19 that smaller weights are assigned to larger coefficient estimates. In other words, these weights (w_j) reduce bias by shrinking large coefficient estimates less, via pilot estimates ($\tilde{\beta}_j$). Therefore, small coefficient estimates are kept under threshold while large ones are asymptotically unbiased. By doing so, the coefficient paths are stabilised and the interpretability and predictability are enhanced.

To respond to the oracle procedure, adaptive lasso aims to pursue large enough non-zero coefficients, while being consistent in feature selection. With a appropriate choice of the penalty (λ_n), i.e. $\frac{\lambda_n}{\sqrt{n}} \to 0$ and $\lambda_n n^{(\frac{\gamma-1}{2})} \to \infty$, Zou showed that adaptive lasso asymptotically enjoys the oracle properties.

$$\hat{\boldsymbol{\beta}}_{alasso} = sign(\tilde{\boldsymbol{\beta}}_j)(|\tilde{\boldsymbol{\beta}}_j| - \frac{\lambda_n}{2|\tilde{\boldsymbol{\beta}}_j|^{\gamma}})_+. \tag{2.20}$$

Similar to lasso in the cases of orthonormal inputs, by solving equation 2.19, equation 2.20 is obtained. This is also known as a two-stage approach that a pilot estimate ($\tilde{\beta}_j$) gives a set of weights (w_j), which is constant across all values of penalty (λ). The `glmnet` package in `R` implements the two-stage approach with the `penalty.factor` argument. In case of $w_j = \begin{cases} 0, & \text{if } \tilde{\beta}_j \neq 0 \\ \infty, & \text{if } \tilde{\beta}_j = 0 \end{cases}$, it is known as lasso-OLS hybrid estimator that lasso and OLS are used as feature selection and coefficient estimation respectively.

An alternative is the pathwise approach that the weights (w_j) vary according to the penalty term (λ_n), i.e. $w_j(\lambda_n) = w(\tilde{\beta}_j(\lambda_n))$. Typically, this pilot estimate ($\tilde{\beta}_j$) is produced from lasso so that the penalty term (λ_n) is consistent between the pilot estimates and the adaptive lasso estimators. In case of $w_j = \begin{cases} 0, & \text{if } \tilde{\beta}_j \neq 0 \\ \infty, & \text{if } \tilde{\beta}_j = 0 \end{cases}$, it is known as the relaxed lasso [21].

2.2.4 Regularised Cox's Proportional Hazards Regression

These Regularised methods can be extended to Cox's proportional hazards model. Cox's proportional hazards model specialises in explaining the relationship between patients' survival and their predictor variables.

Suppose that a survival response is the time to an event, (y_i, δ_i) for i^{th} subject where $i = 1, ..., n$; y_i is a survival time where $\delta_i = 0$ if right censored and $\delta_i = 1$ if died. Denote distinct death

times by $t_1 < ... < t_k$ and d_i deaths at time t_i. Cox's proportional hazard model for the i^{th} subject assumes

$$\lambda_i(t|\boldsymbol{x}_i) = \lambda_0(t)exp(\boldsymbol{\beta}^T\boldsymbol{x}_i), \tag{2.21}$$

where $\lambda_i(t|\boldsymbol{x}_i)$ is a hazard function at survival time t given a $p \times 1$ vector of predictors ($\boldsymbol{x}_i = (x_{1i}, ..., x_{pi})$), $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a $p \times 1$ vector of parameters and $\lambda_0(t)$ is an arbitrary baseline hazard function [22]. Baseline hazard function is the risk of death for the population at particular time interval at the baseline i.e. $\lambda_0(t) = \lambda(t|\boldsymbol{x} = \boldsymbol{0}_p)$. Cox's model is regarded as semi-parametric because this baseline function is non-parametric while $exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ is parametric. To generalise Cox's model, any function $h(\boldsymbol{x}, \boldsymbol{\beta})$ can substitute the exponential function ($exp(\boldsymbol{\beta}^T\boldsymbol{x})$) in equation 2.21.

From equation 2.21, a hazard ratio between subject i and j is

$$HR(t, x_i, x_j) = \frac{\lambda_i(t)}{\lambda_j(t)} = \frac{exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}{exp(\boldsymbol{\beta}^T\boldsymbol{x}_j)} = exp(\boldsymbol{\beta}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)(HR(x_i, x_j) \tag{2.22}$$

Censoring is considered to be non-informative about the vector of coefficients ($\boldsymbol{\beta}$) since that baseline hazard function ($\lambda_0(t)$) can be zero at these time intervals. Therefore, conditional on there have been a particular death at time t_i, the probability of death on the i^{th} subject is

$$\frac{exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}{\sum_{j \in R_r} exp(\boldsymbol{\beta}^T\boldsymbol{x}_j)}, \tag{2.23}$$

where R_r is the set of subjects at risk at time t_r. In equation 2.23, the baseline ($\lambda_0(t)$) does not need to be specified.

Hence, the overall log partial likelihood is

$$\ell(\boldsymbol{\beta}) = log\, L(\boldsymbol{\beta}) = log \prod_{r \in D} \frac{exp(\boldsymbol{\beta}^T\boldsymbol{x}_{jr})}{\sum_{j \in R_r} exp(\boldsymbol{\beta}^T\boldsymbol{x}_j)} \tag{2.24}$$

where D is a set of deaths and j_r is a index of death at time t_r. Given that no tied deaths are in this study and non-informative censoring is assumed, it is reasonable to formulate this partial likelihood. Cox showed that this partial log likelihood can be regarded as maximum likelihood in estimating parameters [22]. Similar to equation 2.6, coefficients in regularised Cox's model are estimated via the following criterion

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left\{ \ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} P(\beta_j) \right\}, \tag{2.25}$$

For each iteration in CV, the loss function is the partial likelihood deviance between the omitted values and predictions based on the modelling set evaluated in the `coxnet` package in `R` [23]. `Coxnet` executes iteration algorithm, via coordinate descent, used in the package `glmnet` for fitting estimated coefficients in penalised Cox regression [23]. The formula for CV error in terms of log partial likelihood deviance given the i^{th} set and λ is

$$\hat{Dev}_i(\lambda) = -2[\ell(\beta_{-i}(\lambda)) - \ell_{-i}(\beta_{-i}(\lambda))], \tag{2.26}$$

where $\ell(\beta_{-i})$ is the log-partial likelihood of the modelling set, β_{-i} is the optimal β of the modelling set. The value of the parameter λ minimizing the deviance ($\hat{Dev}(\lambda) = \sum_{i=1}^{10} \hat{Dev}_i(\lambda)$) is chosen.

2.3 Machine Learning Techniques

Both RF and gradient boosting are regarded as ensemble learning that combines a population of simpler base models in order to construct a prediction model while retaining fairly good interpretability [5]. These techniques can apply to both survival and classification analysis.

SVM and NB are regarded as good classifiers with satisfactory predictive power, yet with poor interpretability [5]. Although NB assumes independence of variables, it performs fairly well in this study. In contrast, SVM often tolerates collinearity well with non-linear kernel.

2.3.1 Random Forest

RF builds a forest of de-correlated decision trees and averages them to reduce the variance [24]. A decision tree recursively undergoes binary space partitioning until a stopping rule, e.g. the pre-assigned minimum number of nodes, is reached. An optimal split subdivides the sample space into two daughter nodes, i.e. binary split, based on a pre-defined metric of homogeneity of the response. Via recursively splitting the sample into sub-samples, decision tree adopts binary decision making, so as to accommodate the non-linearity.

If individual decision tree is allowed to grow sufficiently large, in terms of tree depth, it can capture complex interactions between variables and non-linearity relationship due to its flexibility. As a result, low bias and high variance is expected. Averaging this forest of trees of B bootstrap samples, which is called bagged trees, can cancel the noises among them and benefit the

prediction. Bagging averages the predictions ($\hat{f}^b(x)$) across bootstrap samples

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x), \tag{2.27}$$

where $b = 1,..,B$ is the index of bootstrap sample.

A general RF algorithm is listed as follows [5].

Algorithm for RF

1. For b where $b = 1,...,B$,

   a) Create a bootstrap sample of n rows of the data with replacement from the training set of size n.

   b) Grow a decision tree T_b to the bootstrap sample using recursive partitioning by repeating the following procedure:

      i. Randomly sample pre-assigned p_{RF} ($p_{RF} = \sqrt{p}$ for survival and classification as default) attributes from the p variables

      ii. Test for optimal splits or best attribute among these p_{RF} attributes

      iii. Stop up to minimum node size n_{min}

2. The RF $\{T_b\}_1^B$ is the output.

Regarding survival, log-rank splitting rule is applied for the procedure in the `randomForestSRC` package in `R` [25]. The rule is based on the log-rank test for survival data to maximise the survival difference between two daughter nodes [25–27].

With reference to the notations in section 2.2.4, suppose that a split on value s_j for a variable x_j is proposed, i.e. $\mathbf{L} = \{x_i \leq s_j\}$ and $\mathbf{R} = \{x_j > s_j\}$. The log-rank test for an optimal split on an optimal variable maximise left daughter node ($\mathbf{L} = \{x_i \leq s_j\}$)

$$\text{argmax}\,\mathbf{L}(x_j, s_j) = \frac{\sum_{i=1}^{k}(d_{i,l} - Y_{i,l}\frac{d_i}{Y_i})}{\sqrt{\sum_{i=1}^{k}\frac{Y_{i,l}}{Y_i}(1-\frac{Y_{i,l}}{Y_i})(\frac{Y_i-d_i}{Y_i-1})}}, \tag{2.28}$$

where $d_i = d_{i,l} + d_{i,r}$ denotes the number of subjects who died at t_i;

$Y_i = Y_{i,l} + Y_{i,r}$ for the number of subjects at risk or who died at t_i respectively

$$Y_{i,l} = \#\{t > t_i; x_j \leq s_j\} \; and \; Y_{i,r} = \#\{t > t_i; x_j > s_j\};$$

$d_{i,l}$ and $Y_{i,l}$ represent those for the left daughter node while $d_{i,r}$ and $Y_{i,r}$ for the right.

The optimal split for each p_{RF} variable ($s_j^{'}$) can be obtained, by formulating the log-rank test statistics in equation 2.28, i.e. the larger the statistics, the better the split. Thereafter, by comparing the test statistics for all p_{RF} variable, the optimal variable ($x_j^{'}$) can be found. Furthermore, $|\mathbf{L}(x_j^{'}, s_j^{'})| \geq |\mathbf{L}(x_j, s_j)|$ is sought for each parent node for some values of $(x_j^{'}, s_j^{'})$ up to the terminal node.

The prediction outcome is an ensemble risk score of mortality aggregated from terminal nodes of the survival trees. The higher the score, the higher the risk. This in turns generates Harrell's Concordance index (C-index) for prediction error [28]. This gives the probability of a randomly chosen subject who died of CRC in the validation cohorts had a higher risk than a subject who did not.

Regarding classification, the optimal split rule rests on the least misclassification rate. The prediction for a subject in the validation cohorts is the majority vote of the RF trees, i.e. $\hat{C}_{rf}^{B}(x) = majority\ vote\{\hat{C}_b(x)\}_1^B$ [5].

Variable importance allows the interpretation of the RF model. Based on the specific variable, it is calculated as the improvement in the split-criterion for each split per tree and this is assembled throughout the forest of trees.

Assuming equal variance(σ^2) for each variable, the variance for an average of B i.i.d. variables is $\frac{1}{B}\sigma^2$ and that of identically distributed but unnecessarily independent, which is a feature of bagging, is $(\rho + \frac{1-\rho}{B})\sigma^2$, where ρ is the positive pairwise correlation. This reveals that the variance will shrink to $\rho\sigma^2$ as B increases. Intuitively, the pairwise correlation of trees are controlled by limiting the p_{RF} attributes and further reduce the variance ($\rho\sigma^2$). The training performance is often assessed by the OOB error rate, i.e. the out of the bootstrap sample mean prediction error. This error is misclassification rate for classification and Harrell's C-index using cumulative hazard for survival [25, 29].

2.3.2 Gradient Boosting

Gradient boosting carries out ensemble learning of weak classifiers, mostly as decision trees. Ensemble learning of weak classifiers, also known as "weak learners", tries to reduce bias and variance. It can be divided into two main components, weak classifiers and additive model.

Although these weak learners only perform slightly better than random guessing in terms of error rate, the hypothesis of ensemble learning is to build a strong learner via the unity of these learners. Boosting sequentially creates a set of M weak classifiers $\boldsymbol{f}_m(x)$, where $m = 1,2,...,M$ to repeatedly evolved versions of the dataset [5].

Gradient boosting can be viewed as an algorithm to minimise the loss function with respect to the fitted function $f(x_i)$

$$\hat{\boldsymbol{f}} = \underset{f(x_i)}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathbf{L}(y_i, f(x_i)) \right\}, \tag{2.29}$$

where $\mathbf{L}(y_i, f(x_i))$ is the loss function for $f(x_i)$ and y_i. The loss function is Bernoulli, i.e. logistic regression for classification with 0 and 1 outcomes,

$$\mathbf{L}(y_i, f(x_i)) = 1 + \exp(-2(2y_i - 1)f(x_i)), \tag{2.30}$$

and Cox Proportional Hazard for survival analysis [30].

An additive model is fitted as

$$\boldsymbol{f}_M(x) = \sum_{m=0}^{M} \boldsymbol{\gamma}_m \boldsymbol{h}_m(x) + constant, \tag{2.31}$$

where $\boldsymbol{h}_M(x)$ is the weak learner function obtained from steepest descent step and $\boldsymbol{f}_0 = \boldsymbol{h}_0$ is an initial learner [5,31,32]. Gradient boosting executes an numerical optimization technique named steepest descent, also known as gradient descent, i.e. $\boldsymbol{h}_m = -\rho_m \boldsymbol{g}_m$ where ρ_m is a scalar, also known as step length and the $\boldsymbol{g}_m$ is the gradient. A component of the gradient $\sum_{i=1}^{n} \mathbf{L}(y_i, f(x_i))$ at $\boldsymbol{f} = \boldsymbol{f}_{m-1}$

$$g_{im} = \left[\frac{\partial \mathbf{L}(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x_i)=f_{m-1}(x_i)}. \tag{2.32}$$

Hence, $\boldsymbol{f}_m(x)$ is sequentially induced from $\boldsymbol{f}_{m-1}(x)$ in previous step and so on. It is a weighted sum of weak learners function ($\boldsymbol{h}_m(x)$).

The outline of the algorithm is described as follows [5].

Algorithm for gradient boosting

1. The initial learner is fitted as

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathbf{L}(y_i, \gamma) \right\} \tag{2.33}$$

2. At each step of m,
   the negative gradient of a loss function with respect of $\boldsymbol{f}_{m-1}$ is computed

$$\boldsymbol{f}_m(x) = \boldsymbol{f}_{m-1}(x) + \underset{\boldsymbol{h}_m}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \mathbf{L}(y_i, f(x_i) + \boldsymbol{h}_m(x_i)) \right\}, \tag{2.34}$$

   where $\boldsymbol{h}_m$ is computed from the training set with x_i and g_{im} in equation 2.32.

3. $\boldsymbol{f}_M(x)$ is the output.

It can be seen that the greedily incremental expansion of $\boldsymbol{f}_m$ will end up overfitting the training set in equation 2.34. To avoid overfitting problem for this parsimonious approach, gradient boosting is often executed under the constraints of the complexity of decision trees, as well as CV. Another approach is to apply shrinkage

$$\boldsymbol{f}_m(x) = \boldsymbol{f}_{m-1}(x) + \underset{\boldsymbol{h}_m}{\operatorname{argmin}} \left\{ \upsilon \sum_{i=1}^{n} \mathbf{L}(y_i, f(x_i) + \boldsymbol{h}_m(x_i)) \right\} \tag{2.35}$$

where $0 < \upsilon < 1$ is the learning rate of the boosting procedure and weighs the loss function and weak learner step [5]. This performs similarly as the regularised models seen previously.

2.3.3 Support Vector Machines

SVM constructs an optimal separating kernel for classification problem and is a generalisation of support vector classifier for linear problem. The idea of this classifier comes from maximal margin classifier, which is a solution of the following optimisation problem if a separating hyperplane exists

$$\underset{\beta_0,\ldots,\beta_p,M}{\operatorname{argmax}} \ \mathbf{M} \text{ subject to } \begin{cases} \Sigma_{j=1}^{p} \beta_j^2 = 1, \\ \boldsymbol{y}_i \Sigma_{j=1}^{p} \mathbf{x_j} \beta_j \geq \mathbf{M}, \end{cases} \tag{2.36}$$

where $y_i \in \{-1, 1\}$ is the numerical response for classification and $\mathbf{M}$ is the margin, which is the minimal distance from the training observations to the separating hyperplane [33].

The support vector classifier generalises maximal margin classifier to non-separable cases by

restructuring the optimisation problem as

$$\underset{\beta_0,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n,M}{\operatorname{argmax}} \mathbf{M} \text{ subject to } \begin{cases} \sum_{j=1}^{p} \beta_j^2 = 1, \\ y_i \sum_{j=1}^{p} \mathbf{x_j} \beta_j \geq \mathbf{M}(1-\varepsilon_i), \\ \sum_{i=1}^{n} \varepsilon_i \leq \mathbf{C}, \end{cases} \tag{2.37}$$

where $\varepsilon_i \geq 0$ is the slack variable that tolerates individual observation violating the margin and $\mathbf{C}$ is a non-negative tuning parameter, also known as cost or budget, that bounds the sum of the slack variables [33]. In other words, the budget ($\mathbf{C}$) bounds the amount of error permitted that a larger budget allows a larger margin or mixing zone, also known as soft margin. For illustration of the soft margin in a partition plot, see figure 3.9, section 3.3.2. It shows how this flexibility helps classify the observations. Support vectors refer the observations violating the margin, which is closely related to the budget ($\mathbf{C}$). In bias-variance trade-off, larger budget that gives a lower variance solution tends towards underfitting, and vice versa.

The solution of equation 2.37 can be represented in terms of inner products ($\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle$)

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle, \tag{2.38}$$

where $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \sum_{j=1}^{p} x_{1j} x_{2j}$, $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1p})$ and $0 < \alpha_i < \mathbf{C}$ (see proof B.15, Appendix B) [33].

With regards to classification of non-linear decision boundaries, kernels are applied to allow flexible boundaries between classes. Kernels are functions which quantify the similarity of vectors. The radial basis kernel ($K(\boldsymbol{x}, \boldsymbol{x}_i)$) used inside $f(\boldsymbol{x})$ in this study can be expressed as a generalisation of the inner product

$$K(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}_i\|^2) = \exp(-\gamma \sum_{j=1}^{p} (\boldsymbol{x}_j - \boldsymbol{x}_{ij})^2), \tag{2.39}$$

where ($\|\boldsymbol{x} - \boldsymbol{x}_i\|^2$) is the squared Euclidean distance between two vectors and the parameter γ is a positive constant [33]. Consequently, observations vectors near the evaluation point for $f(x)$ are comparatively influential. This parameter γ changes the decay of radial basis function with distance that larger γ gives a more rapid and more local effect, i.e. complex $f(x)$, and vice versa. Hence, the SVM solution with radial basis kernels can be rewritten as

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\alpha}_i \boldsymbol{y}_i K(\boldsymbol{x}, \boldsymbol{x}_i). \tag{2.40}$$

The optimal budget (**C**) and parameter γ are estimated by CV in this study.

2.3.4 Naïve Bayes

NB is a standard method for classification problem of high dimensional data of which the density estimation is sophisticated. Classifiers based on Bayes theorem can be phrased as

$$P(C_j|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_j)P(C_j)}{P(\boldsymbol{x})}, \tag{2.41}$$

where C_j is the class and $\boldsymbol{x}$ is the matrix of selected variables. The prediction outcome is the highest class probability $P(C_j|\boldsymbol{x})$. Since $P(\boldsymbol{x}|C_j)$ can be expensive to calculate, NB assumes independence for the categorical variable x_j given C_j, i.e. conditional independence

$$P(\boldsymbol{x}|C_j) \approx \prod_{i=1}^{n} P(x_i|C_j). \tag{2.42}$$

Hence, it is regarded as naïve for this optimistic assumption. Under this assumption, the equation 2.41 can be rephrased as

$$\begin{aligned} P(C_j|\boldsymbol{x}) &\propto P(C_j, \boldsymbol{x}) \\ &= P(C_j)P(x_1|C_j)...P(x_n|C_j) \\ &= P(C_j)\prod_{i=1}^{n} P(x_i|C_j). \end{aligned} \tag{2.43}$$

$\boldsymbol{x}$ and $\boldsymbol{y}$ are considered to be evidence in Bayesian terms. This evidence can test the hypothesis that an observation in the validation cohorts belongs to a specific class. The highest probability of each class is the prediction. For continuous variables x_j, the densities from a probability density function is evaluated as Gaussian distribution with mean and variance estimated from the data.

If $P(x_i|C_j) = 0$ for some variables such that $P(\boldsymbol{x}|C_j) = 0$, a Laplacian correction, i.e. a small value or a pseudo-count, is applied to avoid it from happening. Given sufficient sample size, it does not affect much the results.

2.4 Oversampling

Oversampling is a technique to overcome the imbalance of data, which is common in medical data. In this study, 83% of the overall cases were censored and merely 17% patients that died. Given the class imbalance, the outcome is biased towards the censored. For example, without oversampling, a fitted RF model actually predicts all new observations in the validation cohorts as censored in this study.

One method to mitigate the effects of imbalance is to adjust the sample size of the minority class by duplicating the observations of the minority. This traditional method is offered in the `ROSE` package in `R`. SMOTE, which is provided in the `DMwR` package in `R`, produces artificially created data by means of bootstrapping and k-nearest neighbor (k-NN) [34]. Bootstrapping is a method for overcoming the lack of sufficient data and decreasing bias by randomly drawing observations in the dataset with replacement. Bootstrapping is useful for medical studies since the sample sizes are often small with many variables. K-NN algorithm averages the values or takes majority vote of the k-nearest neighbours. SMOTE algorithm first computes the difference between the variable vectors of k-NN via linear interpolation ($\|x - x_i\|$) and then multiplies this difference with a random number between zero and one; a new synthetic observation (x_{new}) is created by adding this adjusted difference into the variable vectors.

$$x_{new} = x + random(0,1) \times \|x - x_i\|, \tag{2.44}$$

where x belongs to the minority class and $i = 1, ..., n_{minority}$ is the number of observations in the minority class.

In this study, k in k-NN is set to 5 closest neighbours, which is typical. Studies have showed that SMOTE can improve accuracy of classification problem [34].

CHAPTER THREE

ANALYSIS AND DISCUSSION

This section gives an overview, analysis and discussion of the data.

For details of the procedures, R script and data files, see list A.1, Appendix A.

3.1 Exploratory Analysis

This study comprises three separate cohorts of stage 1 and 2 CRC patients, with sample sizes of $n_1 = 113$, $n_2 = 56$ and $n_3 = 62$. Two of those ($n_1 = 113$ and $n_2 = 56$) experienced surgical resection hospitals in Edinburgh, in the UK, during 2002-2003 and 2004 respectively. The remaining cohort ($n_3 = 62$) was obtained from the National Defence Medical College Hospital (NDMCH), Japan, dated between 2006 and 2011. The slides of specimens from Edinburgh and Japan were of cross-sectional and longitudinal cut respectively. The follow-up times of the Edinburgh and Japan cohorts were up to 11.5 and 8.6 years respectively. The Edinburgh cohort ($n_1 = 113$) from 2002-2003 served as training set while the other two were used for validation purposes.

The clinical data consist of clinicopathological and spatial variables. The density variables not only quantify the density at different margins, namely, WTS, TBs, Invasive Margin (IM) and Core Tumour (CT), but also formulate the proximity data. The latter explains how certain type of cells surround another at every 50 μm. For the description of variable coding, see tables B.2 and B.1, Appendix B.

3.1.1 Overview Statistics

In the training dataset, the proportion of the censored patients is 5.5 times more than those who died of CRC. Male and female patients are almost equally represented. Patients diagnosed as pT4 or poorly differentiated primary tumour according to the tumour grading rule have a higher proportion of deaths. For details, see figure B.1 and table B.3, Appendix B.

Regarding the continuous variables, the distributions for macrophage and lymphocytic infiltrates are mostly left-skewed. For the response, the median of DSS is 103.44 months and the distribution is right-skewed. This is partly due to the high proportion of the censored patients, i.e. most participants lived beyond the period of follow up, which is up to 11.5 years, i.e. 138 months. Given that this median is high, it is an imbalanced sample.

3.1.2 Correlation

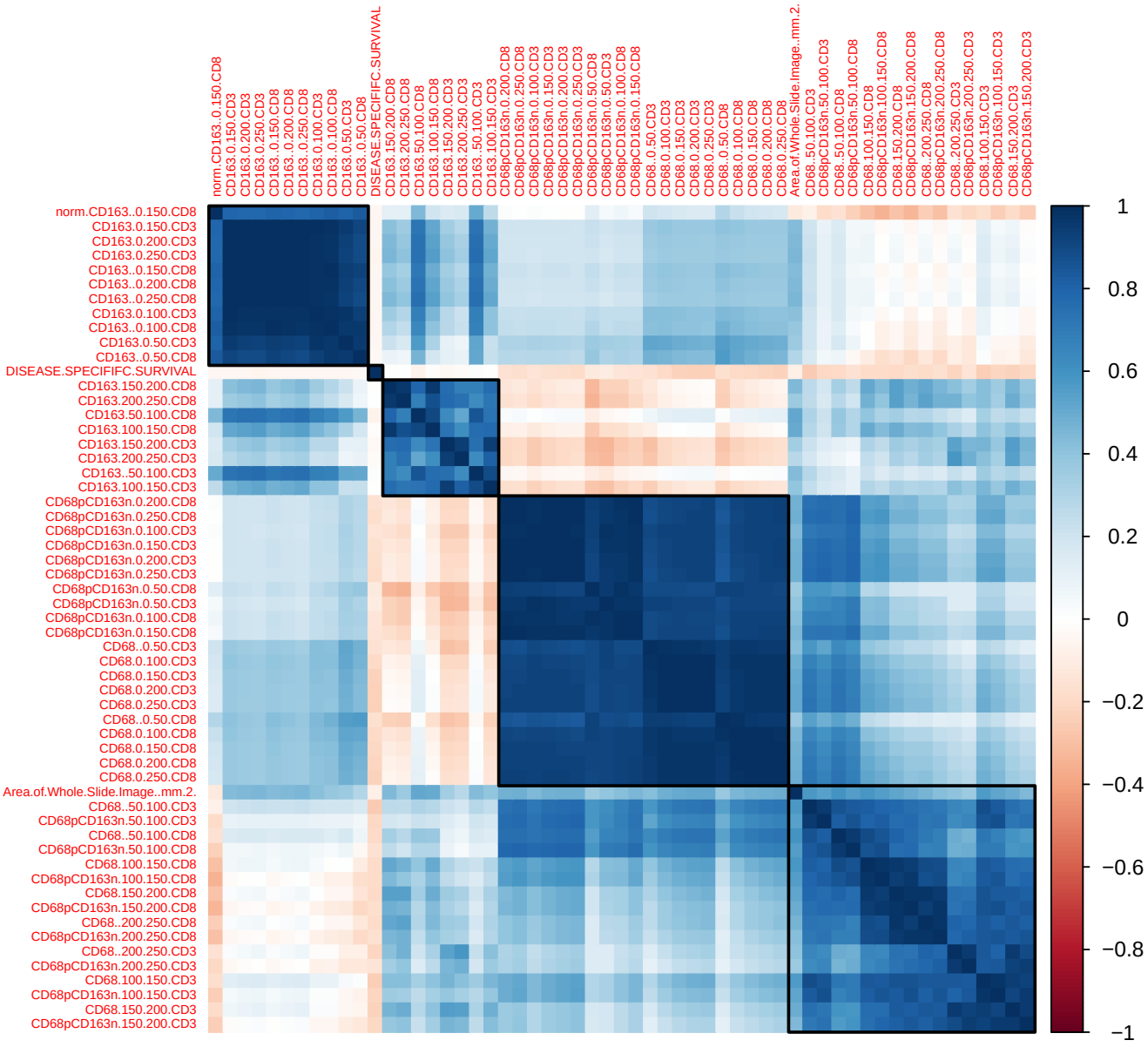


Figure 3.1: Spearman rank correlation matrix between the predictors (x- and y-axis) and Ward's hierarchical clustering. Blue colours represent positive correlations and red colours negative correlations (see legend). Five blocks of high positive correlation are in bold squares.

Figure 3.1 illustrates the non-parametric Spearman rank correlation between the continuous predictors and DSS. Ward's method is used for hierarchical clustering of predictors. This method minimises the increase in total sum of squared correlation distances within the clusters. The five blocks of highly positive correlations identify the grouping of predictors according to their biological and spatial nature , i.e. macrophage infiltrates at distinct locations can be identified.

The third and fourth block are slightly negative correlations; the former denotes M2 phenotype macrophage infiltrates and the latter is M1. Furthermore, the positive correlations among the variables can be seen in these plots. This poses difficulties to distinguish the genuine covariates when modelling.

DSS, the isolated second block, is uncorrelated with the rest. This implies that the response does not have a monotonic relationship with other predictors, i.e. non-linearity.

For outputs of similar results, see figures B.2 ,B.3 and B.4, the latter two using complete average linkage, as a good practice to compare clustering with different methods. Ward's method performs better as expected due to spatial overlaps across clusters.

3.2 Survival Analysis

All the modelling methods in this section were applied on both the original and SMOTEoversampled training dataset. The oversampled dataset has 999 observations ($n_{smote} = 999$) with 504 censored and 495 deaths.

3.2.1 Regularised models

The ridge model is a benchmark model thanks to its holistic view of using all variables. In contrast, the lasso and adaptive lasso ($\gamma = 1, 2, 3$) models are greedy approaches, in particular the latter one, for dimensionality reduction. Elastic net ($\alpha = 0.9$) strikes a balance between predictability and interpretability via grouping effect. The relationships among coefficient paths, CV error and shrinkage are explored for each model.

3.2.1.1 Ridge

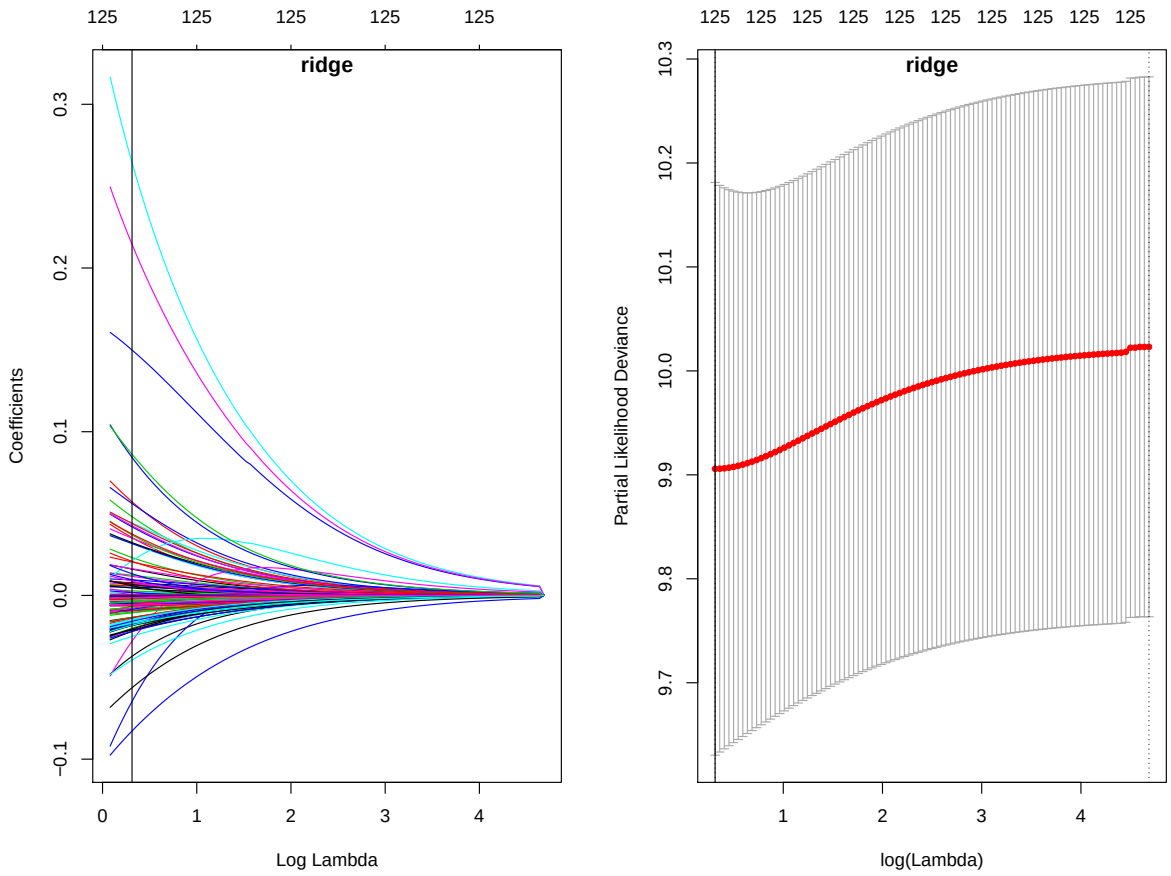


Figure 3.2: Trace (left) and CV (right) plots of ridge regression. The profiles of ridge coefficient (y-axis) paths (different colour for each variable) against $\log \lambda$ (x-axis) (left) and partial likelihood CV deviance (y-axis) (a red dot per iteration) with 95% CI (right) against $\log \lambda$ (x-axis) are shown, where the horizontal axis above denotes the number of selected features; $\log \lambda_{cv.min} = 0.3129576$ at minimum CV (solid vertical lines) and $\lambda_{cv.1se}$ at one standard error (se) from minimum CV(dotted vertical line) are far apart.

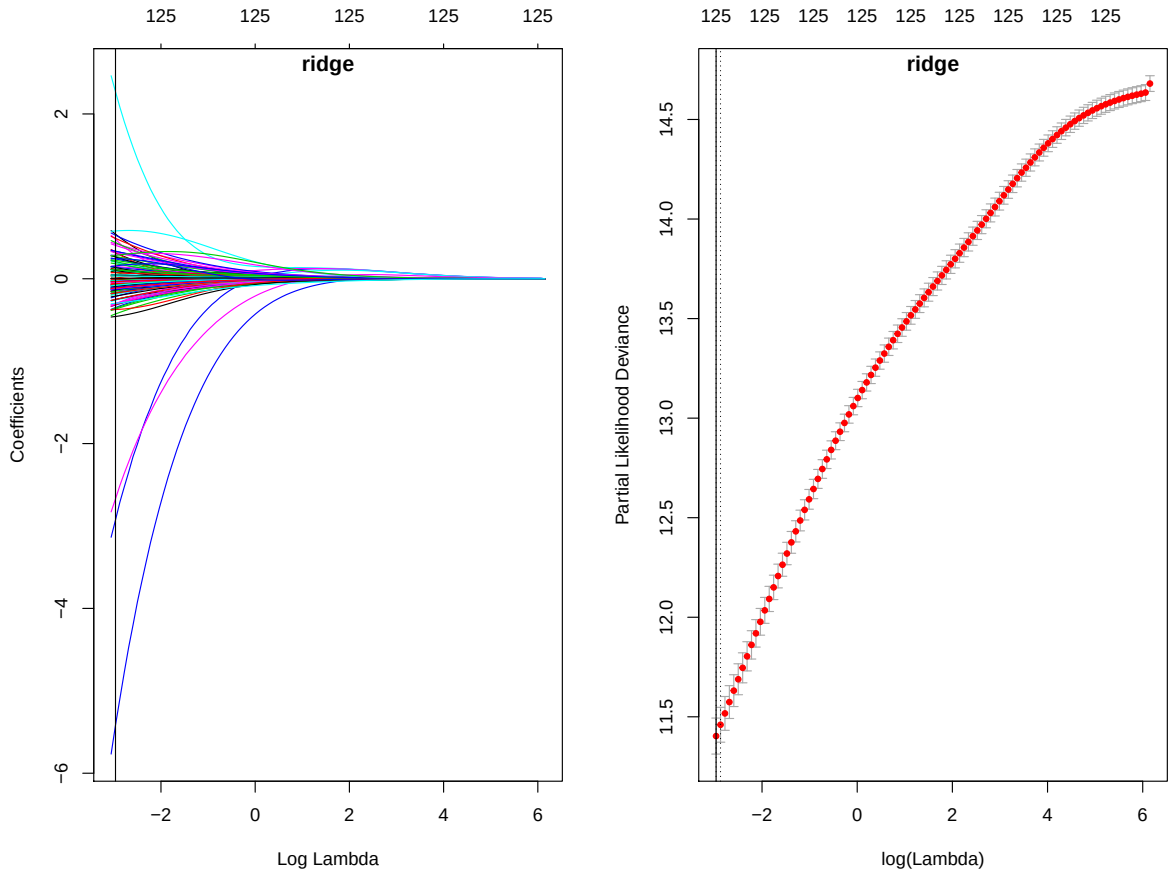


Figure 3.3: Trace (left) and CV (right) plots of ridge regression with SMOTE oversampling. $\log \lambda_{cv.min} = -2.965225$ and closer to $\lambda_{cv.1se}$ but the CI and the gap between $\log \lambda_{cv.min}$ and $\lambda_{cv.1se}$ in the oversampled model is reduced simply because of more observations.

Figures 3.2 and 3.3 (left panel) present the profiles of the coefficient paths for each variables. This shows that the coefficients shrink to almost zero as λ rises. On the right panels of figure 3.2 and 3.3, the partial likelihood CV deviance curve is displayed. It can be seen on the right panel that the CV error increases massively as λ surges. It is because when most variables are shrunk to almost zero together under proportional shrinkage and collinearity, the model cannot explain the data.

Comparing both ridge models (see figures 3.2 and 3.3), the oversampled model has less regularisation. Despite larger oversampled y-axis scale than the original, their coefficients at optimal λ are similar owing to lower regularisation. Indeed, it is found that they return the same models with same coefficients, i.e. oversampling does not affect the ridge.

Among these top five variables (see list A.1, Appendix A), three are continuous and two are

categorical. These continuous variables are all associated with the ratios of CD68 over CD163 in the regions of IM and CT. The categorical variables (ranked below these continuous variables) are differentiation and age. These categorical coefficient weigh less than the continuous ones in terms of magnitude.

3.2.1.2 Lasso

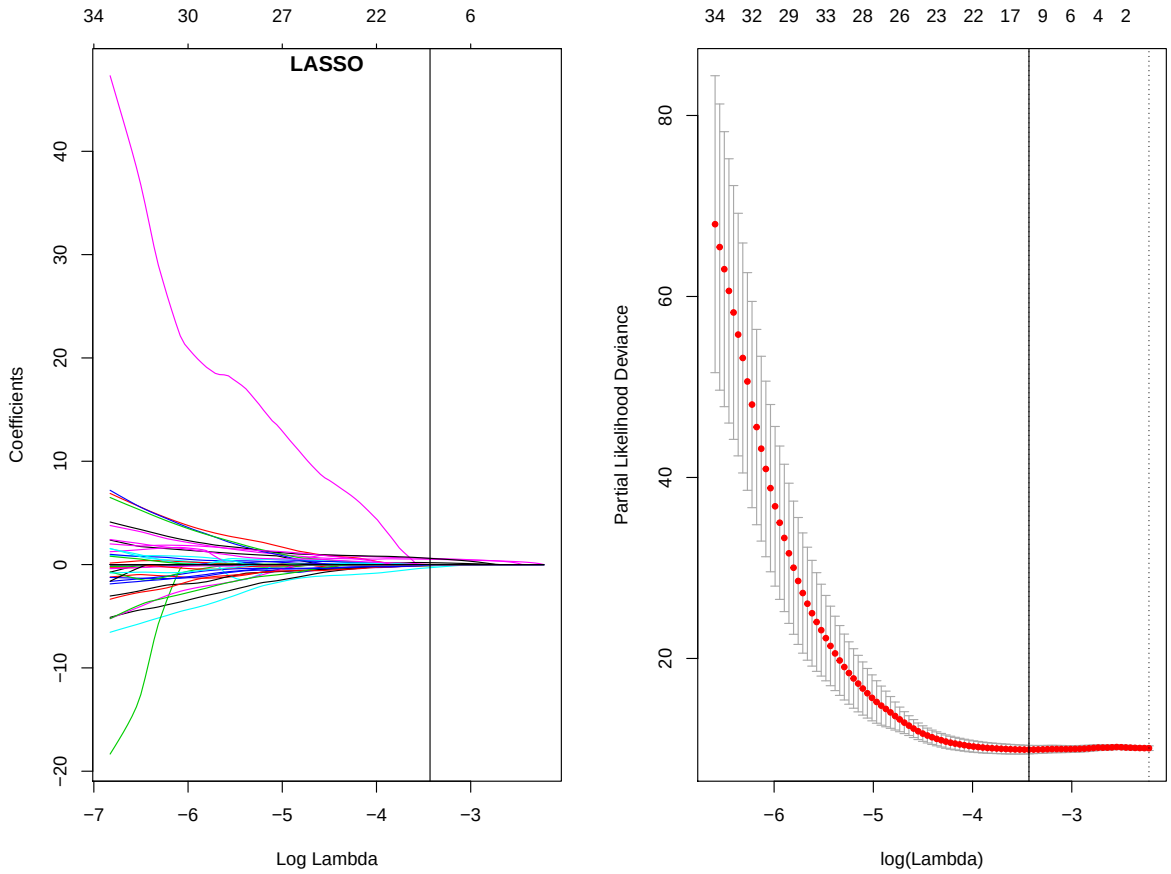


Figure 3.4: Trace (left) and CV (right) plots of lasso regression. $\log \lambda_{cv.min} = -3.43165$, CV and corresponding CI drop as λ increases, i.e. a sign that the model has stabilised.

In figure 3.4, despite a small value of optimal λ, it is sufficient to shrink into eleven selected features in the original lasso model, as shown by the black solid lines on both panels. For details of coefficients, see list A.2, Appendix A. Among these variables, age, the ratio of CD3 over CD8 in CT and the number of CD8 with $0-50\mu m$ of TBs are the distinctive variables of close coefficient value. The higher these variables, the higher the risk of death.

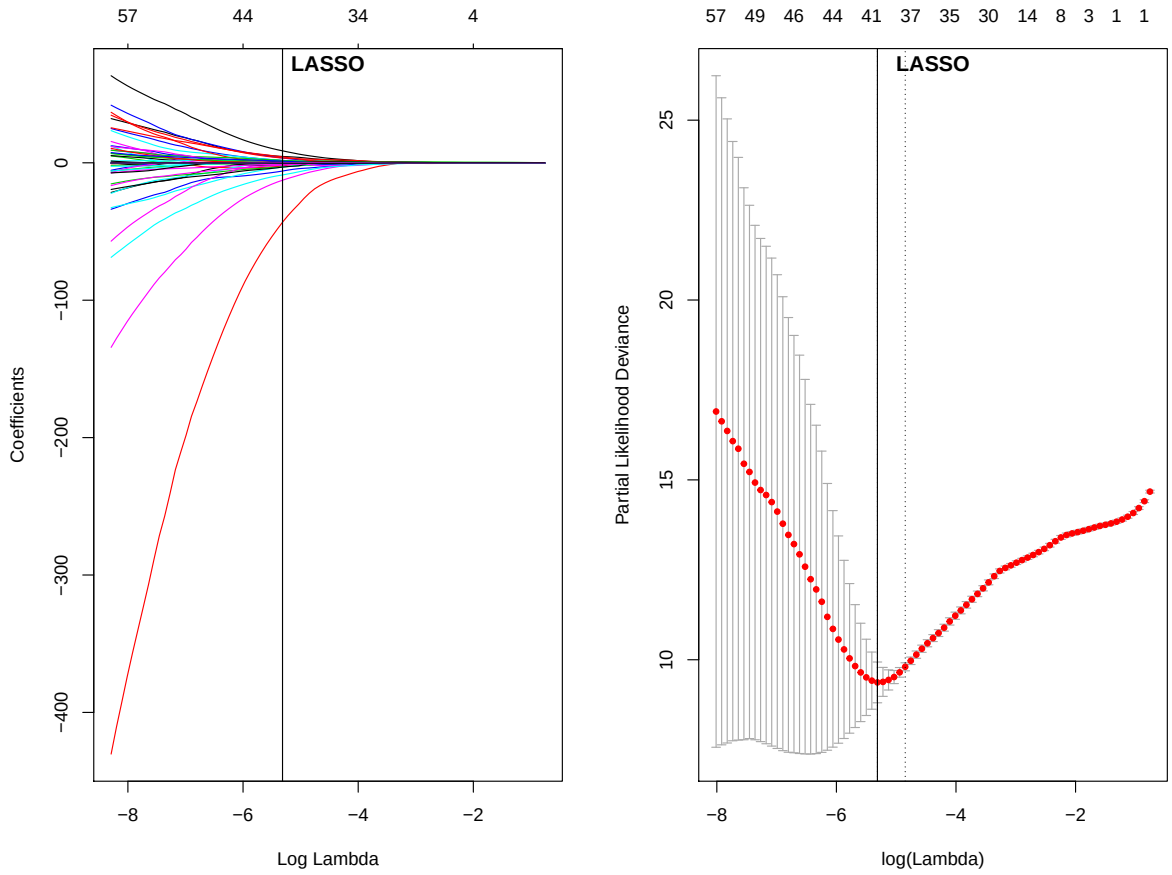


Figure 3.5: Trace (left) and CV (right) plots of lasso regression with SMOTE oversampling. There are 41 selected variables at $\log \lambda_{cv.min} = -5.314327$. The v-shape curve supports the chosen optimal λ as the error surges quickly thereafter.

In figure 3.5, there is more shrinkage than the original model. Also, there are more noises now with more selected variables which makes it harder to interpret (see listing A.2, Appendix A). These influential variables have larger coefficient magnitudes than those in the original lasso model. Apparently, the variable, the ratio of $CD68^{+}CD163^{-}$ over CD163 in IM (the red line on the left panel in figure 3.5), has the largest coefficient value ($\hat{\beta} = -43.24645$) and outweighs other variables. The higher this variable, the lower the risk of death. This may again highlight the characteristics of those dead patients. On the other hand, this model has a lower CV error than the corresponding ridge.

3.2.1.3 Elastic Net

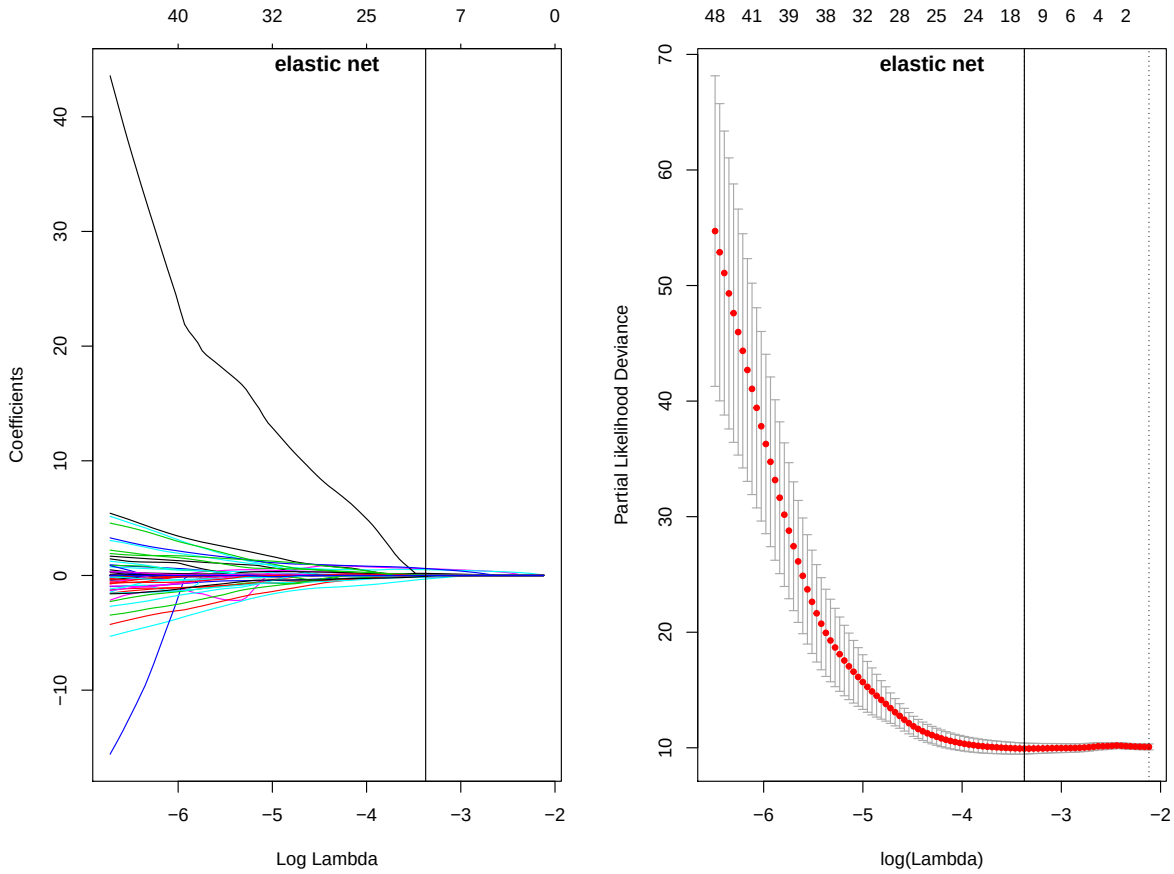


Figure 3.6: Trace (left) and CV (right) plots of elastic net regression ($\alpha = 0.9$). There are 13 selected variables at $\log \lambda_{cv.min} = -3.372807$.

The elastic net parameter was set to 0.9 ($\alpha = 0.9$) because, first, the elastic net empirically performs well if it is close to ridge or lasso models; second, it was hoped to achieve a parsimonious model. In figure 3.6 (left), the profiles of coefficient paths are almost the same as those in the original lasso model as expected.

Not surprisingly, all eleven variables that appeared in the original lasso model are present here with similar coefficient values (see list A.3, Appendix A). Two extra variables included in this model is due to the grouping effect. These extra variables do not seem to improve the deviance at optimal λ compared with the corresponding lasso. Furthermore, their coefficients are negligible compared to the others.

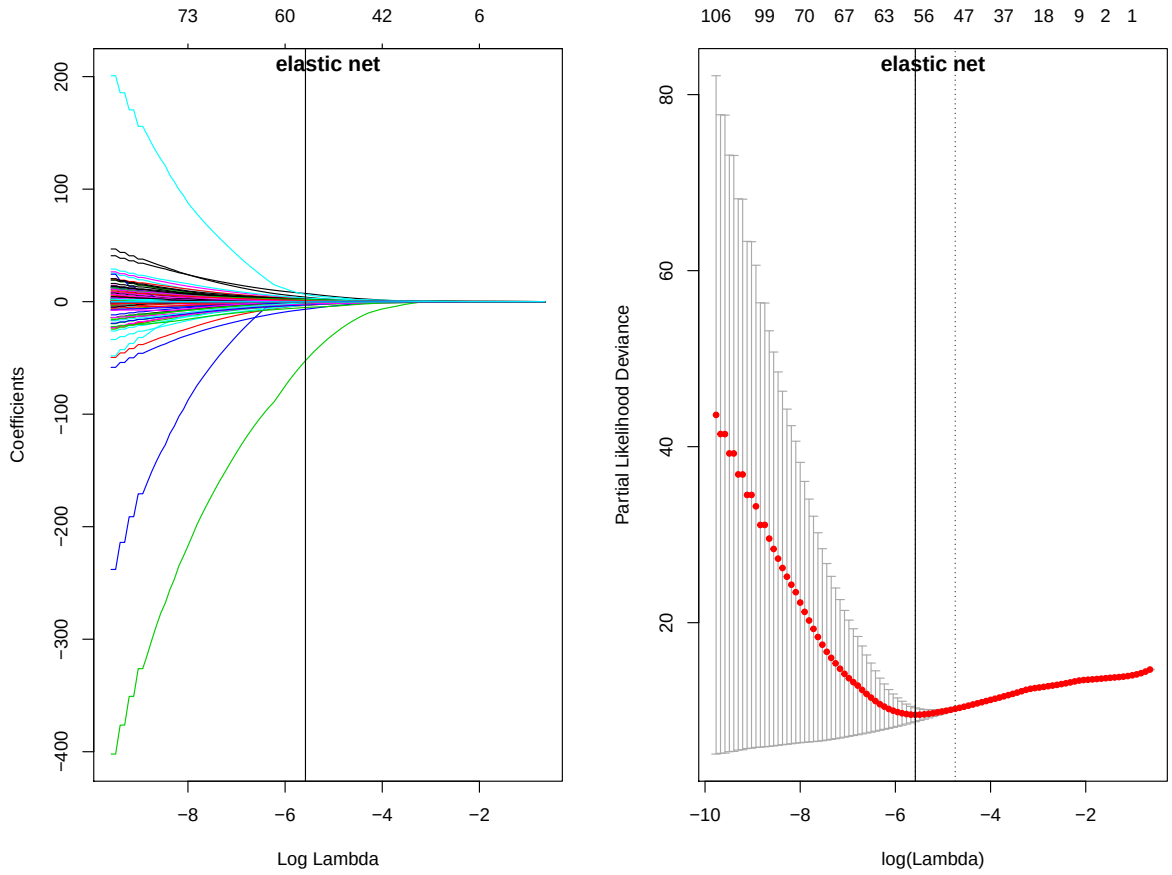


Figure 3.7: Trace (left) and CV (right) plots of elastic net ($\alpha = 0.9$) regression with SMOTE oversampling. There are 58 selected variables at $\log \lambda_{cv.min} = 0.003768414$.

Even more selected variables here make the model harder to interpret. Similar to the oversampled lasso, the most influential variable, which has a coefficient of -52.436740952, is the ratio of $CD68^{+}CD163^{-}$ over CD163 in CT; the rest are all less than 10 in magnitude (see list A.3, Appendix A). Although this CV curve is less drastic than the corresponding lasso, it is still quite certain that the optimal λ should not overfit the data.

3.2.1.4 Adaptive Lasso

To obtain a more consistent and parsimonious model, the oversampled adaptive models were trialled with tuning parameters γ set to be 1, 2 and 3. For the models with the original dataset, the selected variables were all removed when $\gamma = 2$ or 3. The optimal ridge coefficients obtained in subsection 3.2.1.1 were used as the initial estimates ($\tilde{\beta}$).

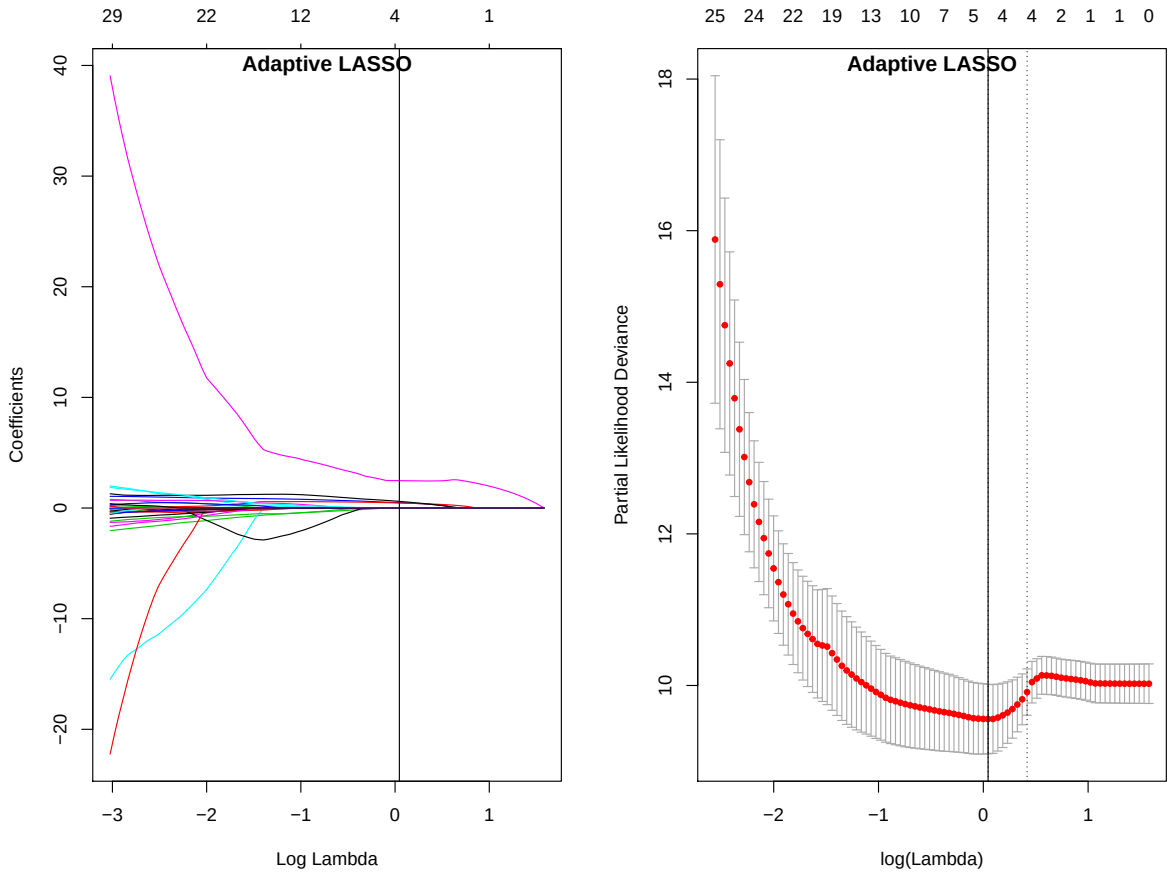


Figure 3.8: Trace (left) and CV (right) plots of adaptive lasso ($\gamma = 1$) regression. Most of the profiles of coefficient paths here are constrained near zero. There are 4 selected variables, the ratio of CD68 over CD163 in CT, age, the ratio of CD3 over CD8 in CT, and CD8 from $0-50\mu m$ of TBs, at $\log \lambda_{cv.min} = 0.04564754$.

Adaptive lasso here can filter out most of the variables. In addition, these variables are consistent at one standard error CV. Therefore, it suggests fthat these variables are stable across most models during CV. Meanwhile, the CV is not very smooth which shows the fluctuations that occur when tuning the parameter λ.

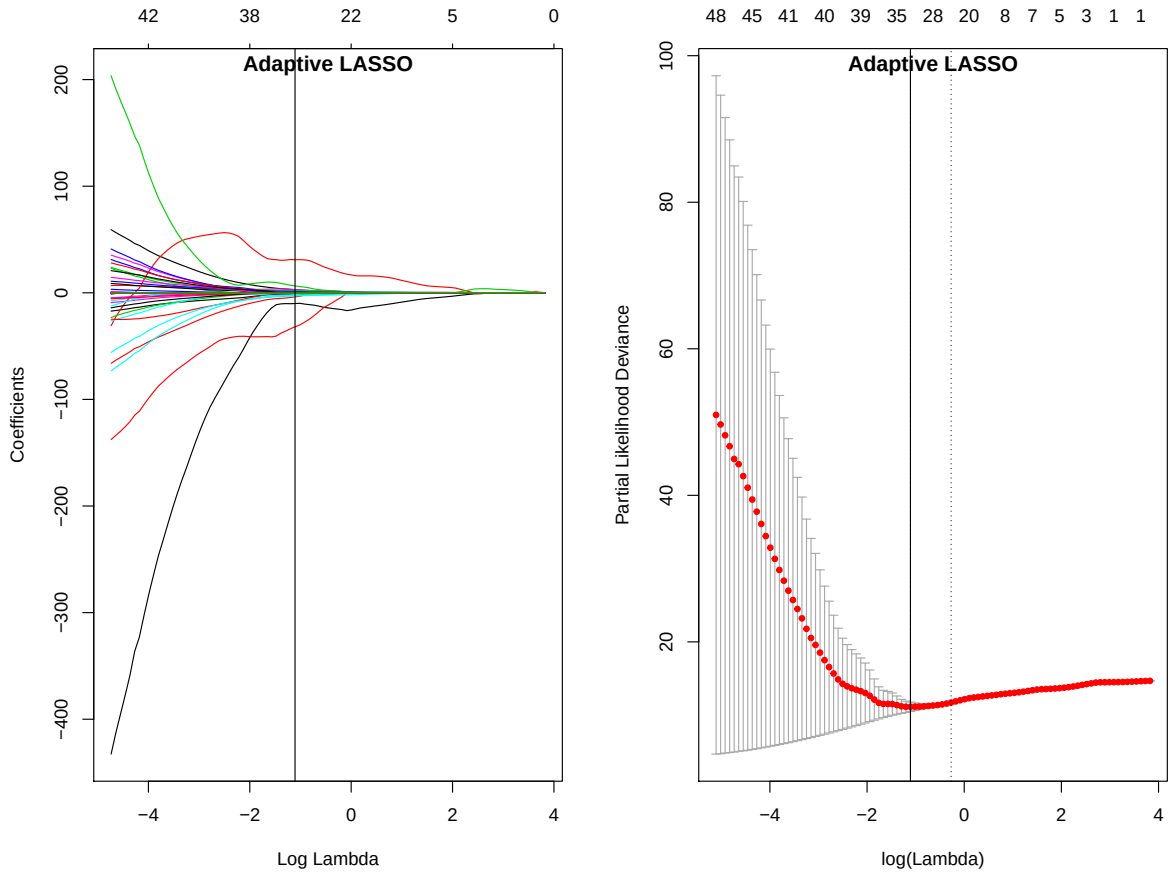


Figure 3.9: Trace (left) and CV (right) plots of adaptive lasso ($\gamma = 1$) regression with SMOTE oversampling. Many profiles of coefficient paths are not restricted near zero with 31 selected variables at $\log \lambda_{cv.min} = -1.107274$(left)

It can be deduced that the tuning parameter (γ) is insufficient to damp the noises generated by SMOTE oversampling as the coefficient paths do not stabilise in figure 3.9 (left). The results for which a heavier tuning parameter (γ) is applied are shown subsequently.

The ratio of CD163 over CD68 in IM and CT and the ratio of CD68 over CD163 in IM are the top two contributing factors, but their coefficients have opposite signs, i.e. the one for the former is -31.869711966, and the latter is 31.037602501. The CV curve has a similar shape to that of elastic net, but elastic net performs slightly better in terms of deviance. This might be owing to the collinearity. This similarity reflects the fact that elastic net and adaptive lasso both try to improve predictability via grouping effect and extra weightings respectively. As a result, both curves are smoother than that of lasso.

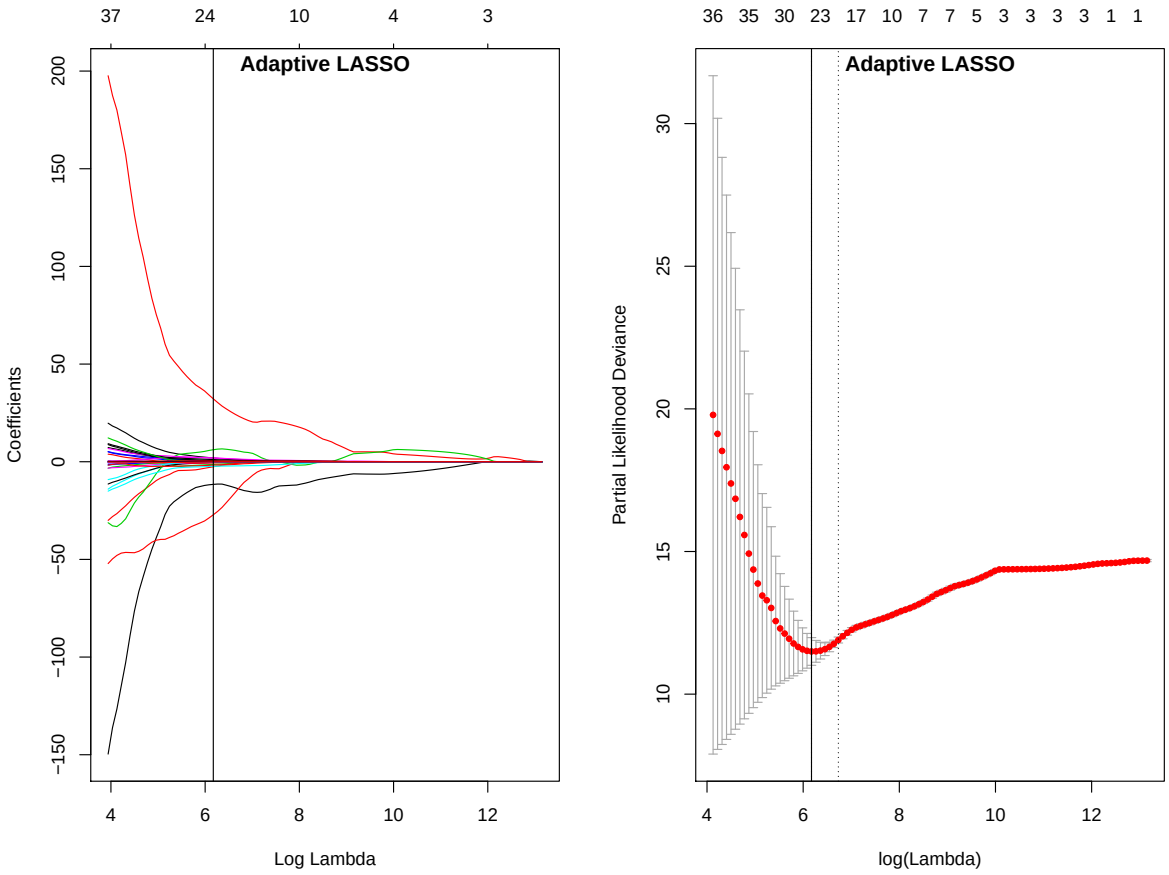


Figure 3.10: Trace (left) and CV (right) plots of adaptive lasso ($\gamma = 2$) regression with SMOTE oversampling. The profiles of coefficient paths (left) are effectively restricted with 24 selected variables at $\log \lambda_{cv.min} = 6.173255$.

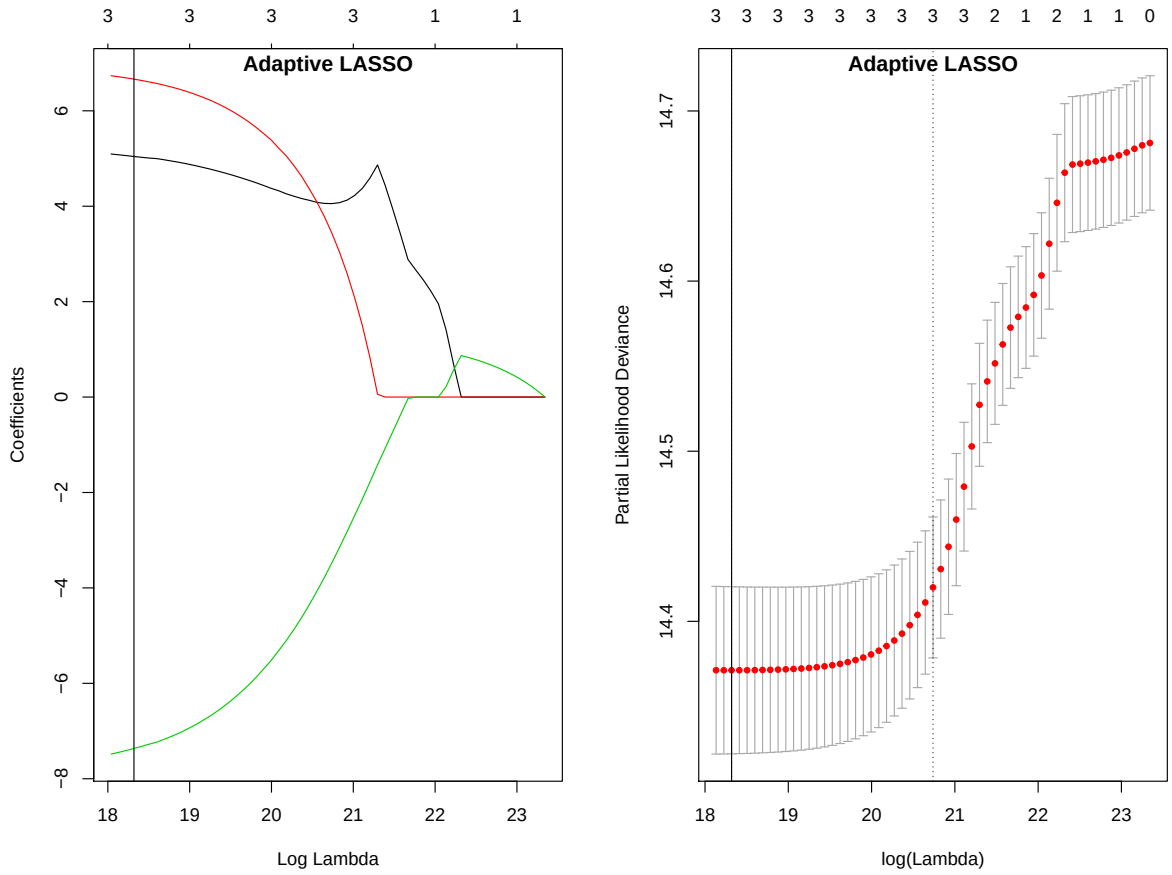


Figure 3.11: Trace (left) and CV (right) plots of adaptive lasso ($\gamma = 3$) regression with SMOTE oversampling. There are only 3 selected variables at $\log \lambda_{cv.min} = 18.31901$, which is comparatively large.

The selected variables are reduced to 24 and 3 for $\gamma = 2$ and 3 respectively (see list A.5). For $\gamma = 2$, the contributing variables are mainly the ratio of CD68 over CD163 in IM ($\hat{\beta} = 32.21094041$) and the ratio of $CD68^{+}CD163^{-}$ over CD163 in IM and CT ($\hat{\beta} = -27.15571444$). With respect to $\gamma = 3$, all the selected variables, which have similar coefficient magnitudes, are the ratio of $CD68^{+}CD163^{-}$ over CD163 in IM, the ratio of CD68 over CD163 in CT and the ratio of CD68 over CD163 in IM. Their coefficients are -7.365046, 6.664329 and 5.043118 respectively.

Generally speaking, the variables associated with macrophage infiltrates are important across these regularised models for both training datasets. Those regularised models fitted from the oversampled dataset tend to emphasise more the variables associated with macrophage infiltrates. In contrast, those fitted from the original dataset have a tendency to include other variables, such as those associated with lymphocytic infiltrates and age group. It is not known if this is driven solely by the dataset or by the inconsistency of regularised models, in particular lasso. Hence, a

comprehensive bootstrap stability study on ridge, lasso and elastic net is described in the next section. Given that adaptive lasso should be consistent due to the extra weights (see subsection 2.2.3.2), it is excluded from the stability study.

Across different types of the regularised models, the variable selections are fairly similar, except where the coefficients may differ.

3.2.1.5 Stability for Regularised Models

The oversampling bootstrap statistics of regularised regressions at optimal λ can be regarded as stability study of the model predictors. For simplicity and being conservative, traditional oversampling is used so that the number of observations for each bootstrap sample is 1000 with class probability of 0.5 ($n_{ovun} = 1000$ and $p = 0.5$) via the `ROSE` package in `R`. The stability on the models fitted with the original and oversampled dataset are compared. It is worth noting that as regularised estimators are biased and thus their CIs generated in this section. Nonetheless, this provides insights to examine their stability and consistency.

3.2.1.6 Ridge

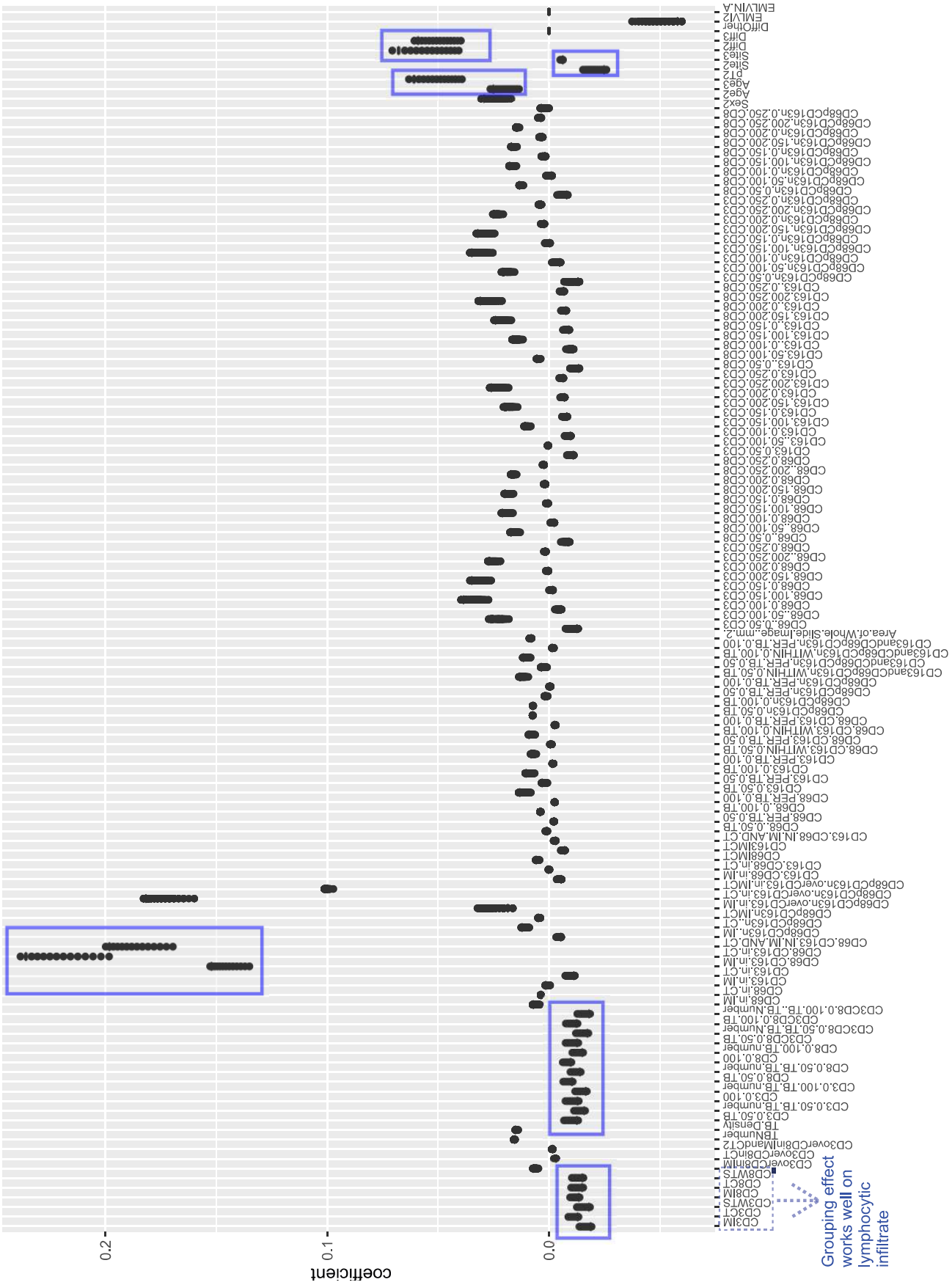


Figure 3.12: Boxplot for ridge regression coefficients (y-axis) at optimal λ with 1000 bootstrap realisations on original training dataset per predictor (x-axis). Grouping effect show coefficients pointing same direction (blue boxes), e.g. a group is from the same lymphocytic category (blue dashed boxes).

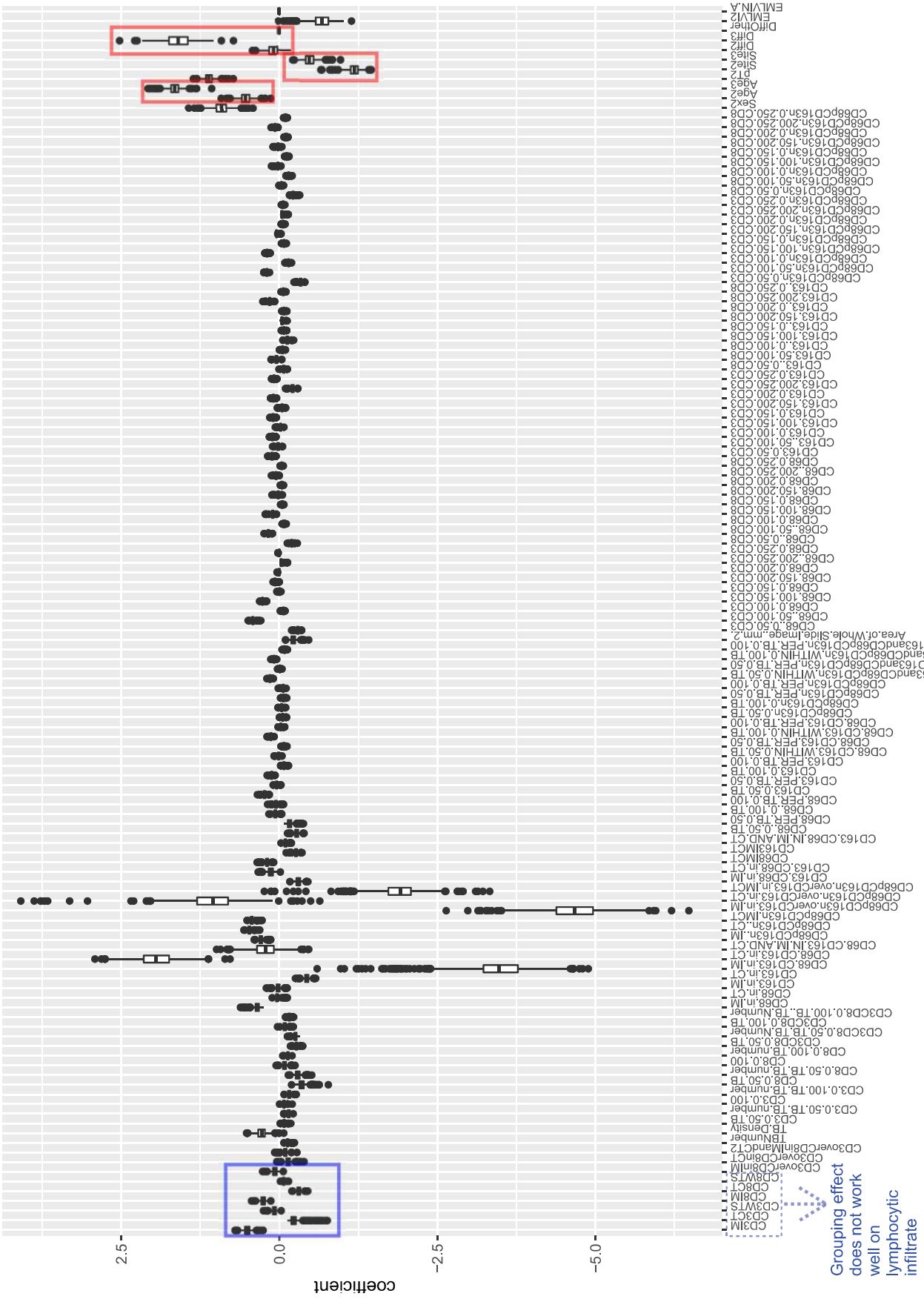


Figure 3.13: Boxplot for ridge regression coefficients (y-axis) at optimal λ with 1000 bootstrap realisations and oversampling n=1000 and p=0.5 per predictor (x-axis). The blue box indicates that grouping effect no longer works well for those lymphocytic infiltrates but still for the categorical variables (red boxes).

As can be seen from figures 3.12 and 3.13, the distributions of ridge-estimated coefficients at the optimal λ ($\hat{\beta}(\hat{\lambda}_{CV.min})$) from both models show signs of the grouping effect. However, the grouping effect does not work very well in the oversampled ridge. Those for the original model are all negative, while those for the oversampled oscillate. This adverse effect is largely because of the increased weighting on the dead patients that gave rise to these undesired noises captured by the model.

For both models, the significant two groups of continuous correlated variables include M1 and M2 phenotype macrophage infiltrates (see table 3.1). For the oversampled ridge, these variables have a notably wide range of 95% CI; while the original model does not have such problem. It can be deduced that these variables in the oversampled ridge fluctuate and compensate for one another in the bootstrap realisations.

Interpreting the coefficients of these continuous variables, it is rather unexpected that both models contradict to one another. The coefficients of these groups in the original model are consistent and on the positive side. The higher these variables, the higher the risk of death. By comparing the coefficients in the oversampled ridge, the M1 and M2 phenotype macrophage infiltrates seem to be the causal variables and thus play a key role within this group. The higher these variables in IM, the lower the risk of death. In contrast, The higher these variables in CT, the higher the risk of death. On the other hand, the ratio of $CD68^{+}CD163^{-}$ over CD163 in IM and CT has a negative impact on the risk of death. It can be deduced that the differences in ratios affect the risk of death.

Most of the categorical variables are significant in the oversampled model; while only pT stage and differentiation are important in the original. These were compared to the baseline (i.e male, age group 1, pT3, left-sided location of primary tumour, moderately differentiated), as shown in table 3.1. In the oversampled model, it suggests that a patient who is female, older, pT4, having left-sided location of primary tumour or well differentiated tumour is more likely to have higher risk of death. In the original model, it indicates that a patient who is pT4 and having poorly or well differentiated primary tumour have higher risk. That said, having a well differentiated tumour with higher risk of death may seem counter-intuitive since this kind of tumour tends to grow and spread slowly. It could be that characteristic was mistakenly identified by ridge regression, or that patients with well differentiated tumours were less active in pursuing treatment.

Original dataset	
Variables	Mean of coefficients [95% CI]
Continuous	
Ratio of CD68 over CD163 in IM	0.15 [0.14,0.15]
Ratio of CD68 over CD163 in CT	0.23 [0.22,0.24]
Ratio of CD68 over CD163 in IM and CT	0.20 [0.18,0.20]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in CT	0.18 [0.17,0.18]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in IM and CT	0.10 [0.10,0.10]
Categorical with baseline	
pT4	0.06 [0.05,0.06]
Poorly differentiated	0.07 [0.05,0.07]
Well differentiated	0.06 [0.05,0.06]
Oversampling dataset	
Variables	Mean of coefficients [95% CI]
Continuous	
Ratio of CD68 over CD163 in IM	−3.45 [-4.33,-2.14]
Ratio of CD68 over CD163 in CT	1.95 [1.38,2.56]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in IM	−4.67 [-5.55,-3.70]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in CT	1.08 [0.34,1.92]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in IM and CT	−1.89 [-2.47,-1.13]
Categorical with baseline	
Female	0.91 [0.63,1.16]
Age2	0.53 [.35,0.71]
Age3	1.65 [1.47,1.84]
pT4	1.11 [0.97,1.24]
Right-sided	−1.19 [-1.35,-1.01]
Rectal	−0.49 [-0.67,-0.31]
Poorly differentiated	0.10 [-0.08,0.29]
Well differentiated	1.60 [1.19,2.05]

Table 3.1: Significant estimated coefficients for ridge regression in bootstrap realisations

3.2.1.7 Lasso

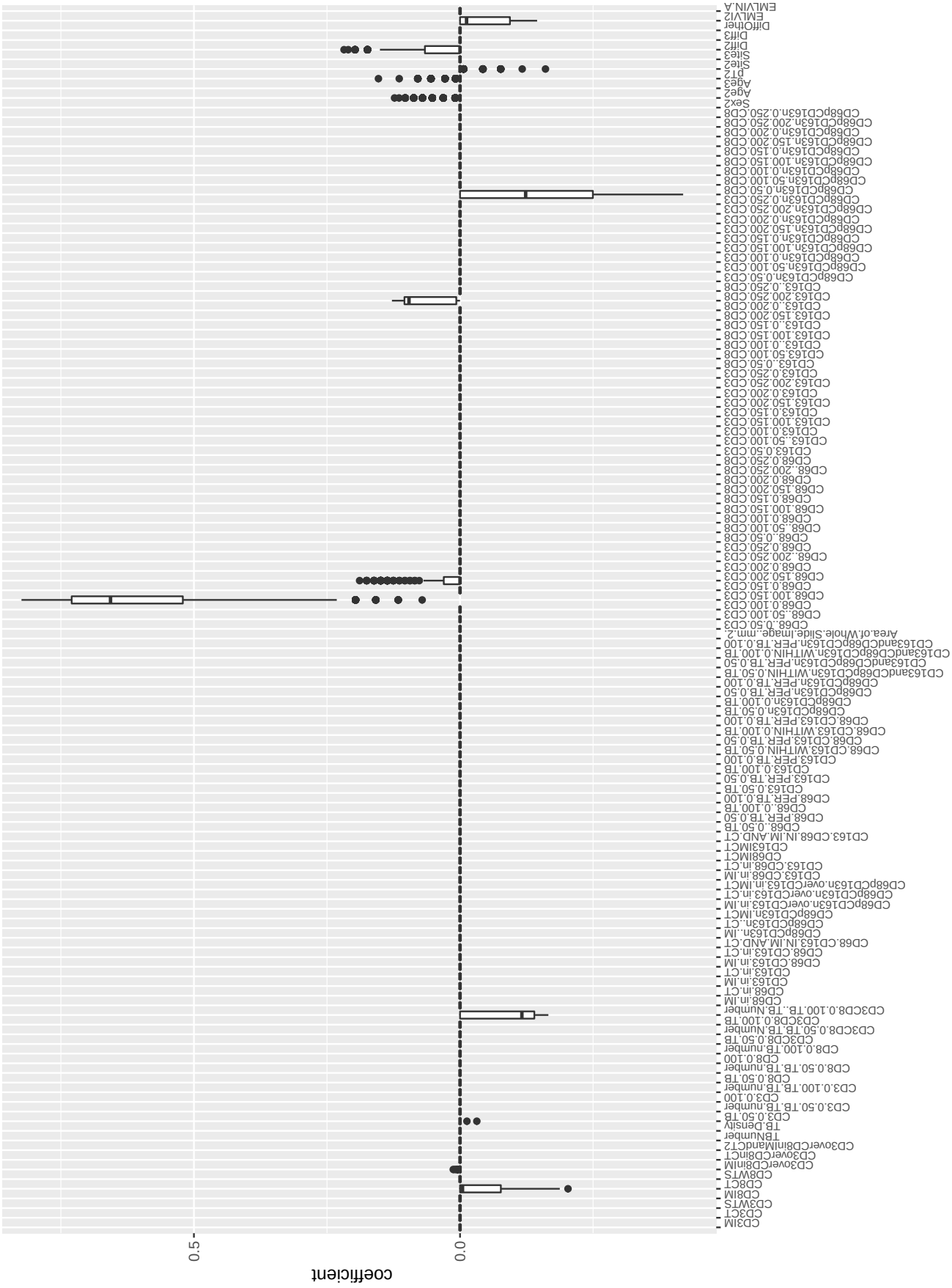


Figure 3.14: Boxplot for lasso regression coefficients (y-axis) at optimal λ with 1000 bootstrap realisations on original training dataset per predictor (x-axis). Lasso effectively shrinks most predictors to zero during bootstrapping.

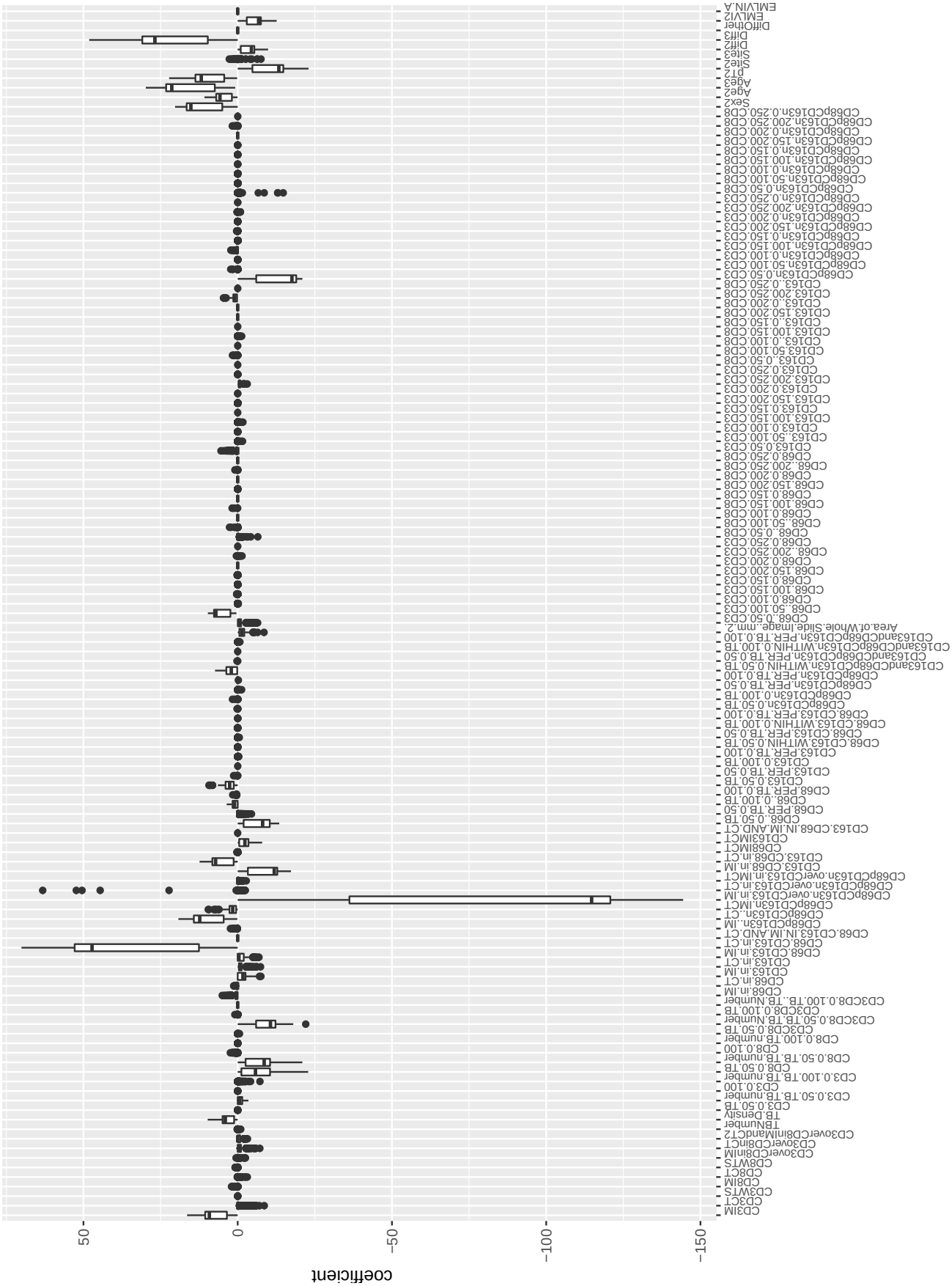


Figure 3.15: Boxplot for lasso regression coefficients (y-axis) at optimal λ with 1000 bootstrap realisations and oversampling n=1000 and p=0.5 per predictor (x-axis). Lasso here cannot effectively shrink predictors to zero during bootstrapping due to the artificial noises of oversampling.

Figure 3.16: Probability of zero (y-axis) for lasso coefficient paths at optimal λ with 1000 bootstrap realisations on original dataset per predictor (x-axis). Although lasso here effectively shrinks most predictors to zero in most iterations, it is still inconsistent across bootstraps.

Figure 3.17: Probability of zero (y-axis) for lasso coefficient paths at optimal λ with 1000 bootstrap realisations and oversampling n=1000 and p=0.5 per predictor (x-axis). lasso here is less consistent in selecting predictors across bootstrap since propensity of being zero for most predictors is not equal to 1.

Figures 3.14 and 3.15 present the sparse solutions of lasso regression ($\hat{\beta}(\hat{\lambda}_{CV.min})$) in bootstrap realisations for both lasso models. Similar to ridge, the coefficients are more stable in the original lasso since most of them are shrunk to zero during bootstrapping.

Since most predictors were shrunk to zero, it is easier to identify the important predictors. On account of feature selection, the correlated variables could no longer aggregate to account for the response. Therefore, these coefficients are markedly larger than the ridge coefficients. On the other hand the coefficients of the oversampled lasso are much larger than those of the original model.

Original dataset	
Variables	Mean of coefficients [95% CI]
Continuous	
CD68 within $100-150\mu m$ of CD3	0.61 [0.20,0.79]
$CD68^{+}CD163^{-}$ within $0-50\mu m$ of CD8	−0.13 [-0.36,0]
CD3CD8 within $100-150\mu m$ of TB per TB Number	−0.08 [-0.16,0]
Oversampled dataset	
Variables	Mean of coefficients [95% CI]
Continuous	
$CD68^{+}CD163^{-}$ in CT	10.30 [1.01,16.76]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in IM	−86.98 [-128.02,-0.87]
$CD68^{+}CD163^{-}$ within 0-50 μm of CD3	−13.62 [-20.09,-0.35]
Categorical with baseline	
Female	11.90 [0.54,18.57]
Age2	4.84 [0.50,8.86]
Age3	16.93 [2.05,25.36]
pT4	0.09 [0.89,15.85]
Right-sided	−10.69 [-16.62,-0.72]
Rectal	−0.09 [-0.82,0.53]
Poorly differentiated	−3.52 [-6.68,0]
Well differentiated	22.40 [1.64,37.37]

Table 3.2: Significant estimated coefficients for lasso regression in bootstrap realisations

The significant coefficients are extracted in table 3.2. For the variables selection of continuous variables, the results from the original and oversampled lasso are dissimilar. CD68 within $100-150\mu m$ of CD3 is the most influential variable for original lasso and this variable does

not appear to be important in the original ridge models. For the oversampled model, it is the ratio of $CD68^{+}CD163^{-}$ over CD163 in IM. For the original model, it highlights the connection between macrophage and lymphocytic infiltrates. By contrast, the oversampled model underlines the macrophage infiltrates in IM.

In contrast to the grouping effect in ridge regression, there are only three significant continuous variables in both models and they are from distinctive groups. In addition, their 95% CIs are substantially wider than the those from corresponding ridge models due to the erratic process of feature selections across bootstrap realisations. This indicates how highly correlated the variables are.

Interestingly, the oversampled lasso highlights the proximity of macrophage infiltrates ($CD68^{+}CD163^{-}$) around lymphocytic infiltrates(CD3) within $0-50\mu m$ as a crucial variable, in contrast to the oversampled ridge. This variable can be suppressed by noises in ridge regression. Its mean in table 3.2and relatively narrow 95% CI in figure 3.15 suggest that this variable is quite stable. With respect to the significant categorical variables, the result of the oversampled lasso mostly agrees with those of the oversampled ridge, except that the sizes of coefficients vary to a fairly large extent. Regarding the original model, the categorical variables are much less important than the continuous variables.

Lasso variables	
Original dataset	
CD68 within $100-150\mu m$ of CD3	
Oversampled dataset	
Ratio of $CD68^{+}CD163^{-}$ in CT	CD68 50-100 μm CD3
Age2	Age3
pT4	CD3CD8 0-50 μm TB/TB Number
Right-sided	Well differentiated
Ratio of $CD68^{+}CD163^{-}$ 0-50 μm CD3	Female
TB Density	Ratio of $CD68^{+}CD163^{-}$ over CD163 in IM
Area of Whole Slide Image mm^2	CD3IM
CD68 within 0-50 μm of TB	CD163 within 0-50 μm of TB
Ratio of CD68 over CD163 in CT	CD163/CD68 in IM
No EMLVI	CD163 in CT
CD8 0-50 μm TB/TB number	
Ranked in descending order according to proportion of appearing times,	(from top to bottom, left to right)[1]

Table 3.3: Variables which appeared in more than 95% lasso bootstrap realisations

The proportion of times that the variables were shrunk to zero in the bootstrap realisations of

lasso are set out in figures 3.16 and lasso, there are twenty-one lasso variables that appeared in more than 95% of the bootstrap realisations listed in table 3.3; while there is only one variable for the original lasso. Nonetheless, this allows assessment of their significance and stability from another perspective. This suggests greater consistency in the oversampled lasso model, as compared to the original model.

It is worth noting that all the significant variables in table 3.2 were present in more than 95% of the bootstrap realisations. This means that these variables are not only significant in terms of coefficients, but also are consistent across bootstrap realisations (see list A.6, Appendix A).

3.2.1.8 Elastic Net

Figure 3.18: Boxplot for elastic net regression coefficients (y-axis) at optimal λ with 1000 bootstrap realisations on original training dataset per predictor (x-axis). The grouping effect in elastic net works well (blue boxes) and sufficiently shrink most predictors to zero.

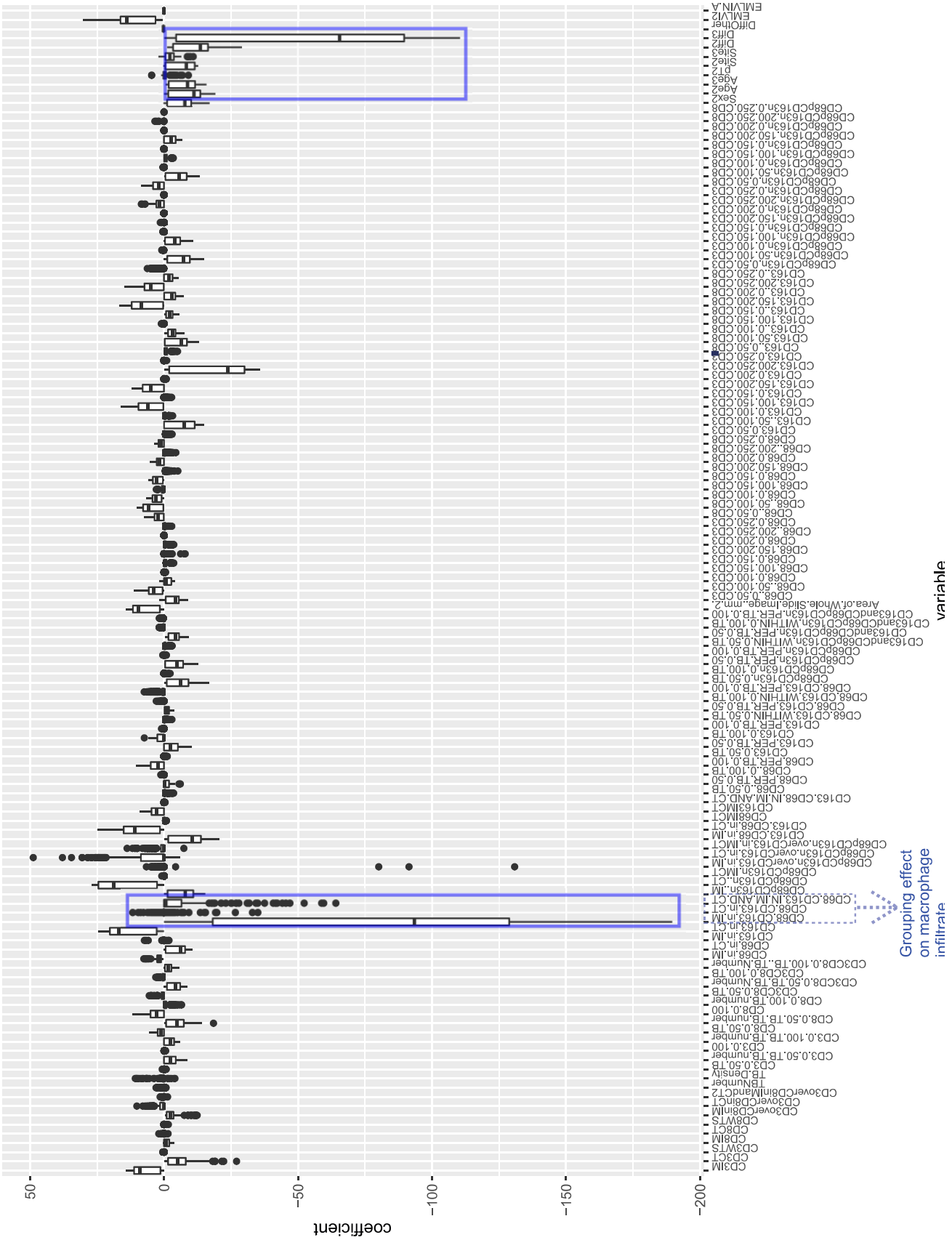


Figure 3.19: Boxplot for elastic net regression ($\alpha = 0.9$) coefficients (y-axis) at optimal λ with 1000 bootstrap realisations and oversampling n=1000 and p=0.5 per predictor (x-axis). Grouping effect here does not as efficiently (blue boxes) as there are many fluctuations.

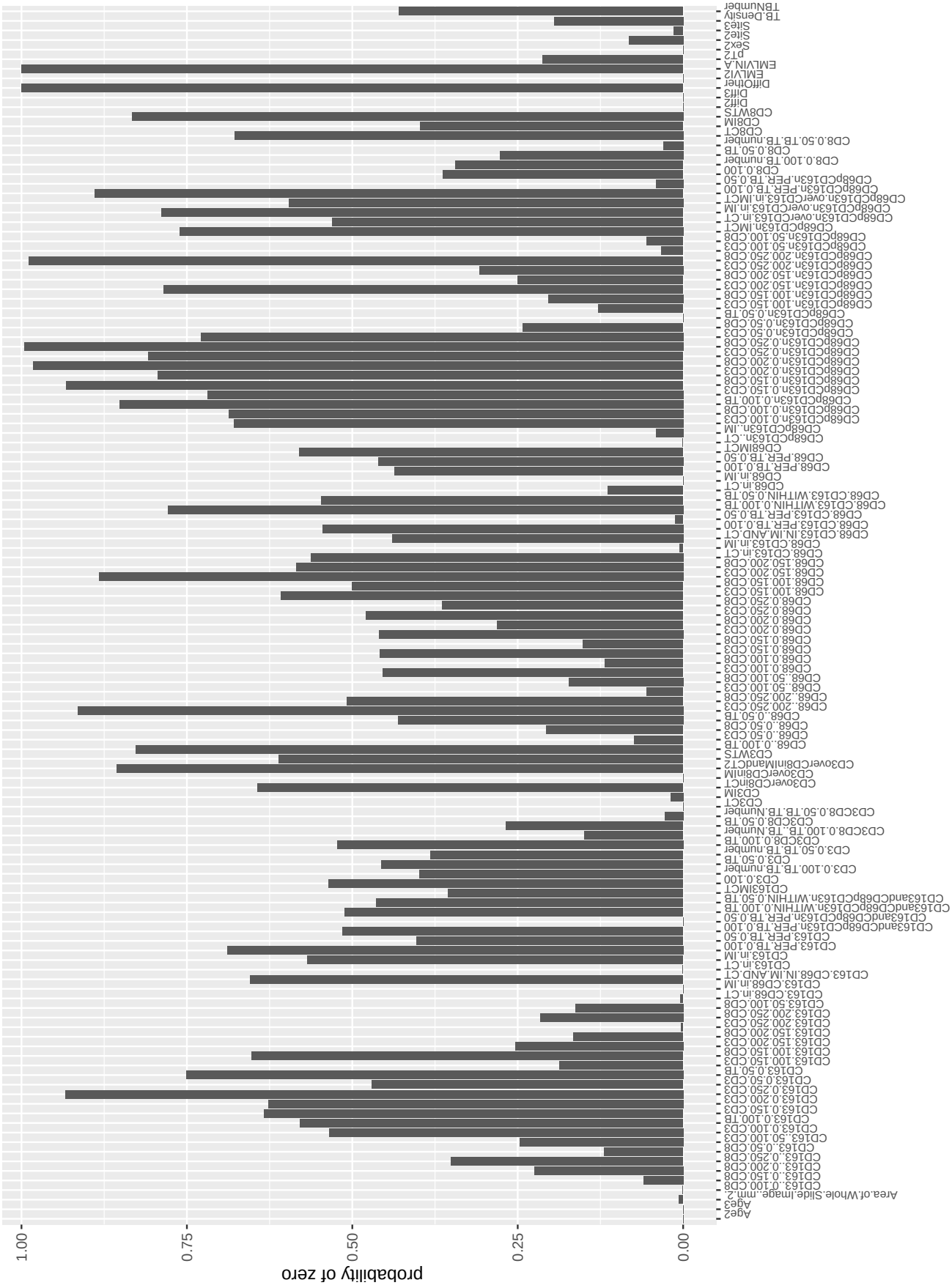


Figure 3.20: Probability of zero (y-axis) for elastic net ($\alpha = 0.9$) coefficients paths at optimal λ with 1000 bootstrap realisations on original dataset. There is less propensity of being zero for most predictors here. It is unknown if it is consistent across bootstraps.

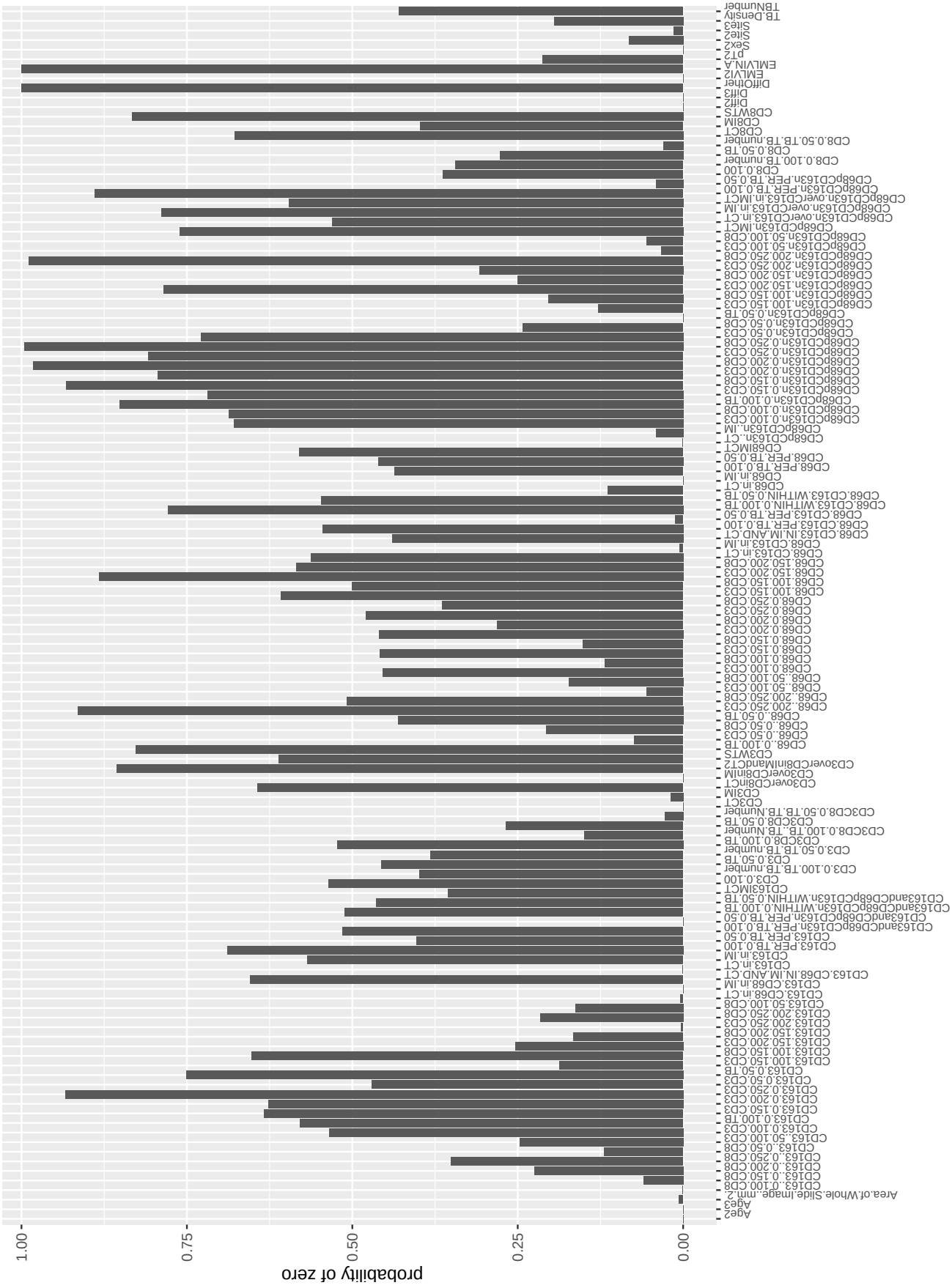


Figure 3.21: Probability of zero (y-axis) for elastic net ($\alpha = 0.9$) coefficients paths at optimal λ with 1000 bootstrap realisations and oversampling n=1000 and p=0.5 per predictor (x-axis). This plot is very similar to figure 3.20 that there are a great deal of fluctuations across bootstraps.

Figures 3.18 to 3.21 show how elastic net ($\alpha = 0.9$) underwent variable selection with grouping effect, i.e. in between ridge and lasso regressions. The grouping effect from the original model is more identifiable than that from the oversampled. The original model shrinks many variables to zero.

Regarding continuous selected variables, the original elastic net selects only two significant variables (see table 3.4). Among these, the 95% CI of $CD68^{+}CD163^{-}$ within $0-50\mu m$ of CD8 touches zero. For the oversampled elastic net, four variables related to the macrophage infiltrates are significant, similar to ridge regression. However, two of these 95% CIs (ratio of CD68 over CD163 in IM and ratio of $CD68^{+}CD163^{-}$ over CD163 in IM and CT) touch zero. These suggest that these variables are not consistent across bootstrap realisations of the corresponding models.

Regarding the proportion of times being shrunk to zero (see list A.7, Appendix A), there are one and forty-four predictors appeared in 95% of the bootstrap realisations for the original and oversampled elastic net respectively. It appears that grouping effect does not help stabilise the selected variables in the original model.

For the oversampled elastic net, table 3.4 suggests that the two continuous variables, ratio of CD68 over CD163 in CT and ratio of $CD68^{+}CD163^{-}$ over CD163 in IM, are important. This is again resulted in between the ridge and lasso models. Regards to categorical variables, they are similar to ridge and lasso regressions.

To summarise ridge, lasso and elastic net models, variables associated with macrophage infiltrates play a key role since these variables are consistent across bootstrap realisations for both original and oversampled datasets. These also show that the fitted models obtained in section 3.2.1 are stable and consistent. On the other hand, it is not surprising that lasso and elastic net ($\alpha = 0.9$) give similar feature selections. Notwithstanding the consistency and stability of these models, the prediction results in the next section can better reflect how their performances are on unseen validation datasets.

Original dataset	
Variables	Mean of coefficients [95% CI]
Continuous	
CD68 within $100-150\mu m$ of CD3	0.52 [0.17,0.75]
$CD68^{+}CD163^{-}$ within $0-50\mu m$ of CD8	−0.09 [-0.29,0]
Oversampled dataset	
Variables	Mean of coefficients [95% CI]
Continuous	
Ratio of CD68 over CD163 in IM	−23.82 [-39.64,0]
Ratio of CD68 over CD163 in CT	32.80 [0.44,50.80]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in IM	−53.89 [-79.68,-3.50]
Ratio of $CD68^{+}CD163^{-}$ over CD163 in IM and CT	−19.97 [-31.91,0]
Categorical with baseline	
Female	9.79 [0.81,14.00]
Age2	5.07 [0.61,7.64]
Age3	14.34 [2.05,20.26]
pT4	10.50 [1.05,15.23]
Right-sided	−9.87 [-14.04,-1.08]
Rectal	0.42 [-0.88,1.76]
Poorly differentiated	−3.32 [-5.77,0]
Well differentiated	14.49 [1.99,2.14]

Table 3.4: Significant estimated coefficients for elastic net regression in bootstrap realisations

3.2.1.9 Predictions

	Validation(EDI)				Validation(JAP)			
	Ridge	Lasso	Elastic net ($\alpha=0.9$)	Alasso ($\gamma=1$)	Ridge	Lasso	Elastic net ($\alpha=0.9$)	Alasso ($\gamma=1$)
C-index	0.4188912	0.3531828	0.4271047	0.4229979	0.6877193	0.7473684	0.7596491	0.6017544
D_{xy}	-0.1622177	-0.2936345	-0.1457906	-0.1540041	0.3754386	0.4947368	0.5192982	0.2035088
S.D.	0.2084399	0.1764507	0.1837074	0.1624047	0.1264497	0.1541257	0.1335654	0.1400909
n	56	56	56	56	61	61	61	61
missing	0	0	0	0	0	0	0	0
uncensored	11	11	11	11	13	13	13	13
Relevant Pairs	974	974	974	974	1140	1140	1140	1140
Concordant	408	344	416	412	784	852	866	686
Uncertain	2,106	2,106	2,106	2,106	2,520	2,520	2,520	2,520

Table 3.5: Validation results for the regularised models fitted from the original dataset. The validation cohorts used here are the Edinburgh cohort after 2004 and Japanese cohort.

	Validation(EDI)					
	Ridge	Lasso	Elastic net ($\alpha = 0.9$)	Alasso ($\gamma = 1$)	Alasso ($\gamma = 2$)	Alasso ($\gamma = 3$)
C-index	0.5400411	0.4722793	0.3634497	0.4414784	0.4435318	0.5462012
D_{xy}	0.08008214	-0.05544148	-0.27310062	-0.11704312	-0.11293634	0.09240246
S.D.	0.1763848	0.1818006	0.1678574	0.1907445	0.1863333	0.1791976
n	56	56	56	56	56	56
missing	0	0	0	0	0	0
uncensored	11	11	11	11	11	11
Relevant Pairs	974	974	974	974	974	974
Concordant	526	460	354	430	432	532
Uncertain	2,106	2,106	2,106	2,106	2,106	2,106

	Validation(JAP)					
	Ridge	Lasso	Elastic net ($\alpha = 0.9$)	Alasso ($\gamma = 1$)	Alasso ($\gamma = 2$)	Alasso ($\gamma = 3$)
C-index	0.4719298	0.477193	0.745614	0.4842105	0.4929825	0.5596491
D_{xy}	-0.05614035	-0.04561404	0.49122807	-0.03157895	-0.01403509	0.11929825
S.D.	0.156966	0.1618898	0.1467068	0.164996	0.1741696	0.1561112
n	61	61	61	61	61	61
missing	0	0	0	0	0	0
uncensored	13	13	13	13	13	13
Relevant Pairs	1140	1140	1140	1140	1140	1140
Concordant	538	544	850	552	562	638
Uncertain	2,520	2,520	2,520	2,520	2,520	2,520

Table 3.6: Validation results for the regularised models fitted from the SMOTE oversampled dataset. The validation cohorts used here are the Edinburgh cohort after 2004 and Japanese cohort.

Any model with a C-index larger than 0.5 is better than random guessing, i.e. this model predicts higher survival chance for observations with longer survival time and vice versa for C-index less than 0.5. The C-index, which is derived from the Wilcoxon-Mann-Whitney two-sample rank test, formulates the AUROC and offers a comparable measure of prediction performance across models [28]. Hence, a C-index of 0.5 corresponds to the AUROC of random guessing classification. Generally, any survival models with C-index around 0.8 are regarded as usable. This index takes into account of all possible pairs of subjects whose survival times can be correctly ordered and reflects the proportion of pairs that are correctly ordered. D_{xy} refers to the Somers' D_{xy} rank correlation, which can be formulated as

$$D_{xy} = 2(c - 0.5) \tag{3.1}$$

where c is the concordance index [28]. If D_{xy} is zero, the model performs the same as random guessing. If D_{xy} is one, it suggests perfect discrimination. S.D. refers to the standard error of D_{xy}. n is the number of observations. There is no missing in the validation cohorts. Uncensored is the number of patients who died. Concordant is the number of concordant pairs between the predictions and observations. Uncertain is the number of pairs of observations which

classification of concordance cannot be obtained due to censoring.

The validation results for the regularised models set out in tables 3.5 and 3.6 are not satisfactory. Elastic net fitted from the original dataset, which performs the best here, has contrasting performance between these two cohorts. It is very likely because of the two cohorts itself. However, why exactly a model fitted from the Edinburgh cohort in 2002-2003 performs better in the Japanese than the Edinburgh after 2004 is not known.

According to the C-index, the SMOTE oversampling does not improve prediction performance. These models generally are outperformed by the corresponding original models. Notwithstanding the undesired performances, the predictability of adaptive lasso improves as γ increases. It is partly because the insufficient observations of dead patients in the training dataset. Another reason could be that the noises induced by SMOTE obscure the underlying signal.

Overall, these models do not achieve a satisfactory result. They produce polarised outputs depending on the validation cohorts. Those predicting disappointingly for the Edinburgh cohort tend to work well in the Japanese. This contrast can be due to the natures of the validation cohorts. In short, these models do not generalise for validation. Strictly speaking, the elastic net fitted from the outperforms the rest of the regularised models.

It is suspected that the reason for these disappointing performances is due to the non-linearity of the relationship. Hence, machine learning techniques are applied in the next section to accommodate this complexity.

3.2.2 Machine Learning Techniques

In this section, RF and gbm models were fitted for the original training dataset. Oversampling is usually not required for these two techniques since they can accommodate the imbalance internally through the bagging and boosting algorithm and the imbalance of the dataset here is not very extreme. Nonetheless, modelling with oversampling was tested but the performance was worse than without oversampling. Therefore, the results are not presented. A comparison of prediction performance between the RF and gbm optimal models is presented at the end of this section.

3.2.2.1 Random Forest

The following combination of parameters setting of the RF survival model was trained with OOB error rate in the `randomForestSRC` package in `R`:

- The numbers of tree built were 5,000 and 10,000;

- The block size was 10, i.e. the cumulative error rate computed for every 10^{th} trees;
- The forest terminal node sizes were 5, 10, 15 and 20;
- The number of pre-assigned p_{RF} as ($\sqrt{p} = \sqrt{124} \approx 11$) and 20.

The optimal model was with 5,000 trees, terminal node size of 5 and $p_{RF} = \sqrt{p}$. The corresponding tree cumulative OOB error rate is 15.23% (see list A.8, Appendix A). In other words, for more complex RF models with larger number of trees, terminal node size and p_{RF}, they perform no better than the simpler. Once the OOB error rate, the RF training can stop. In figure 3.22, this error stabilises at around 5,000 trees. Hence, further addition of trees cannot improve the error rate.

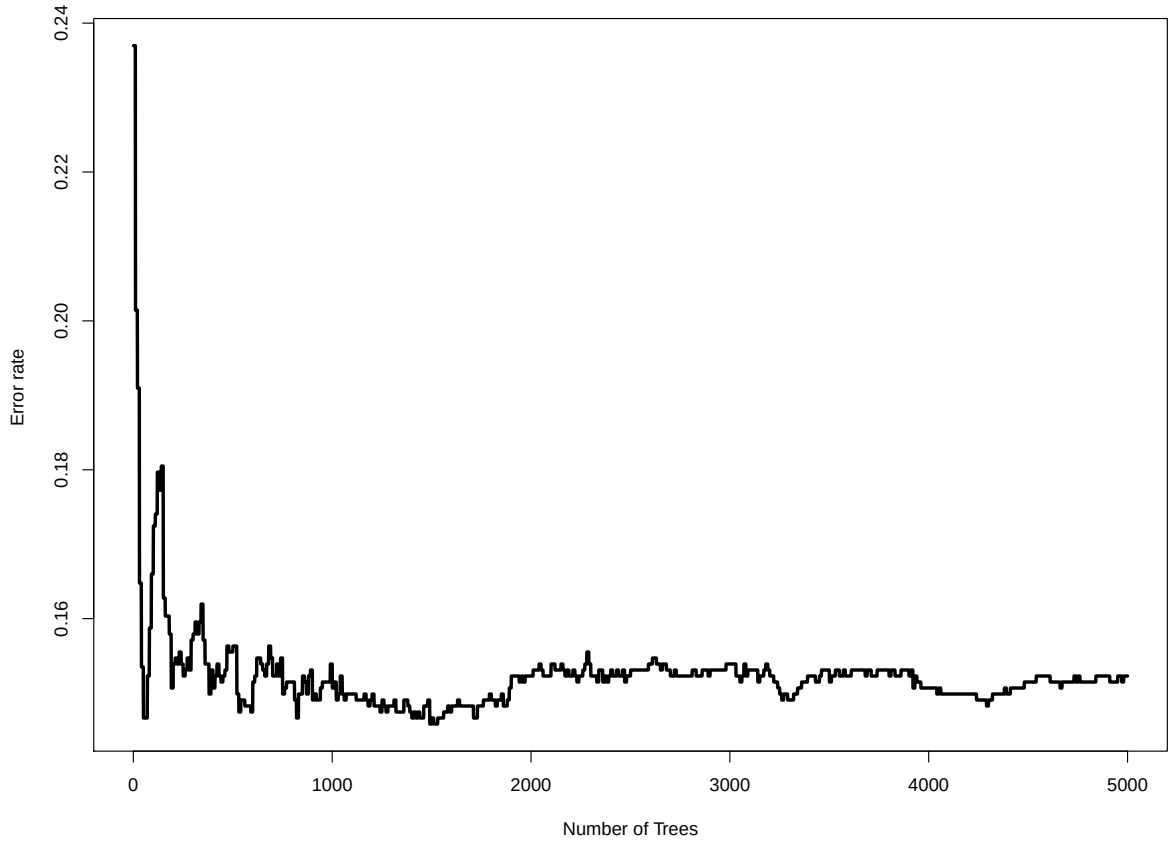


Figure 3.22: OOB error (y-axis) versus number of trees (x-axis) plot for RF model trained with 5,000 trees, block size of 10 and terminal node size of 5. The error rate oscillates up to roughly 4000 trees and stabilises thereafter.

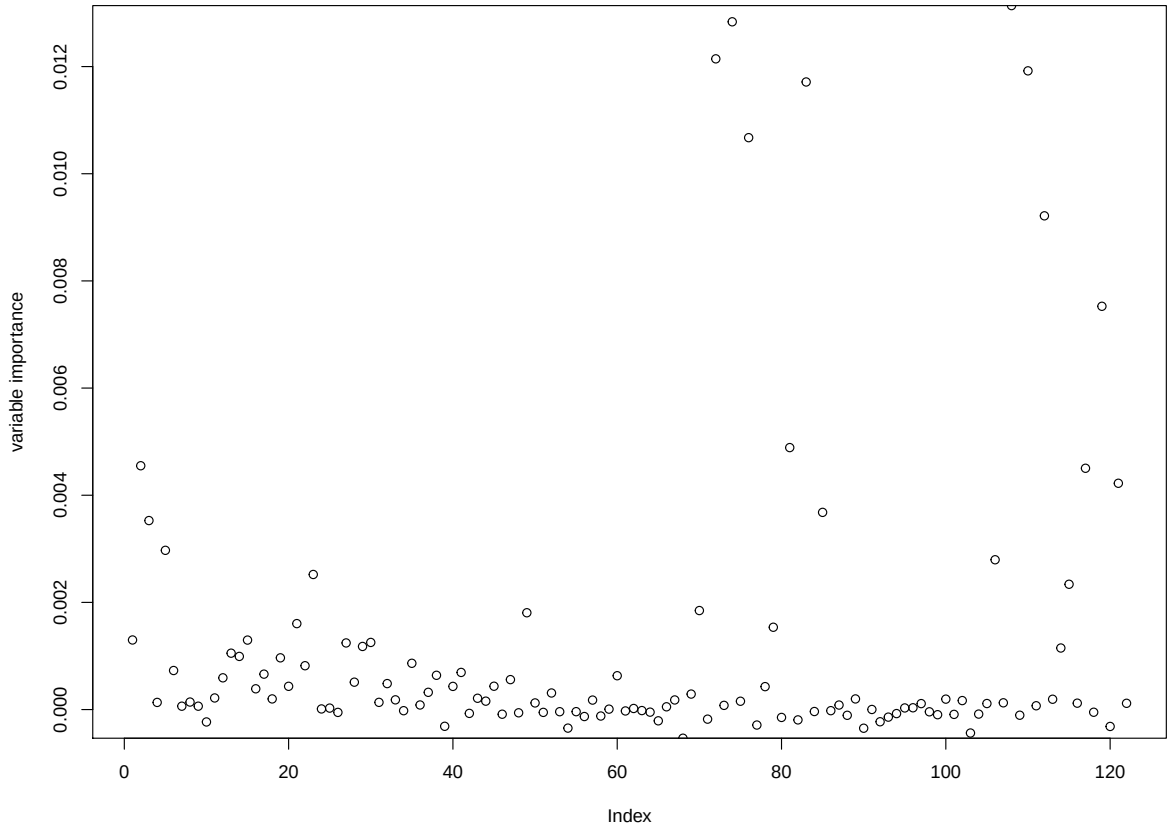


Figure 3.23: Variable importance (y-axis) versus predictor index (x-axis plot for RF model trained with 5,000 trees, block size of 10 and terminal node size of 5. Most points of insignificant predictors scatter around zero and those of the significant are distanced from the rest.

The variable importance in the RF model is summarised according to their index of each variable in figure 3.23. Given that most of the points scatter near zero, these vital variables show a markedly different pattern. This suggests that most of the variables can be safely ignored. Interestingly, the top twenty variables are all proximity data of macrophage infiltrates around lymphocytic infiltrates, as shown in the next figure 3.24. This highlights the significance of the interaction between these infiltrates.

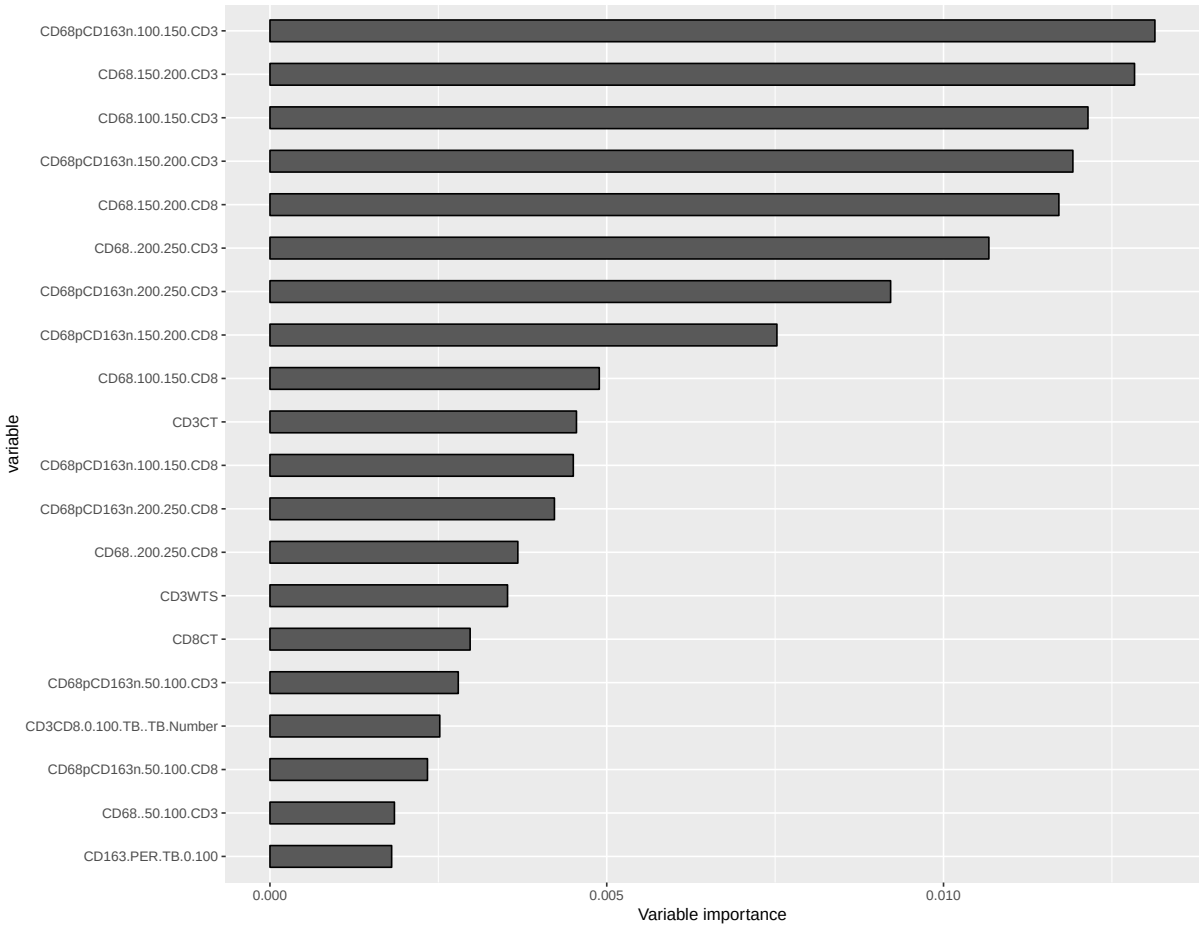


Figure 3.24: Variable importance (x-axis) plot for top 20 variables (y-axis) of RF model trained with 5,000 trees, block size of 10 and terminal node size of 5. There is a sudden drop after the 8^{th} variable, CD68pCD163n.150.200.CD8.

The top twenty variable importance in the RF model is set out in figure 3.24. It is again remarkable that most of the crucial variables here are associated with macrophage infiltrates and their linkage to lymphocytic infiltrates. Specifically, after the top eight important variables, there is an observable cutoff in terms of variable importance. These variables are then followed by the density of lymphocytic infiltrates in CT or WTS. For further details, see list A.8, Appendix A.

In general, these findings mostly agree with the previous regularised models on the importance of the aforementioned interaction. Compared to the regularised models, this RF model further places the importance on the relationship between macrophage infiltrates and survival. Nonetheless, those regularised models, except ridge, often select one among the correlated variables and suppress the rest. As a result, those models had relatively less emphasis on these suppressed variables.

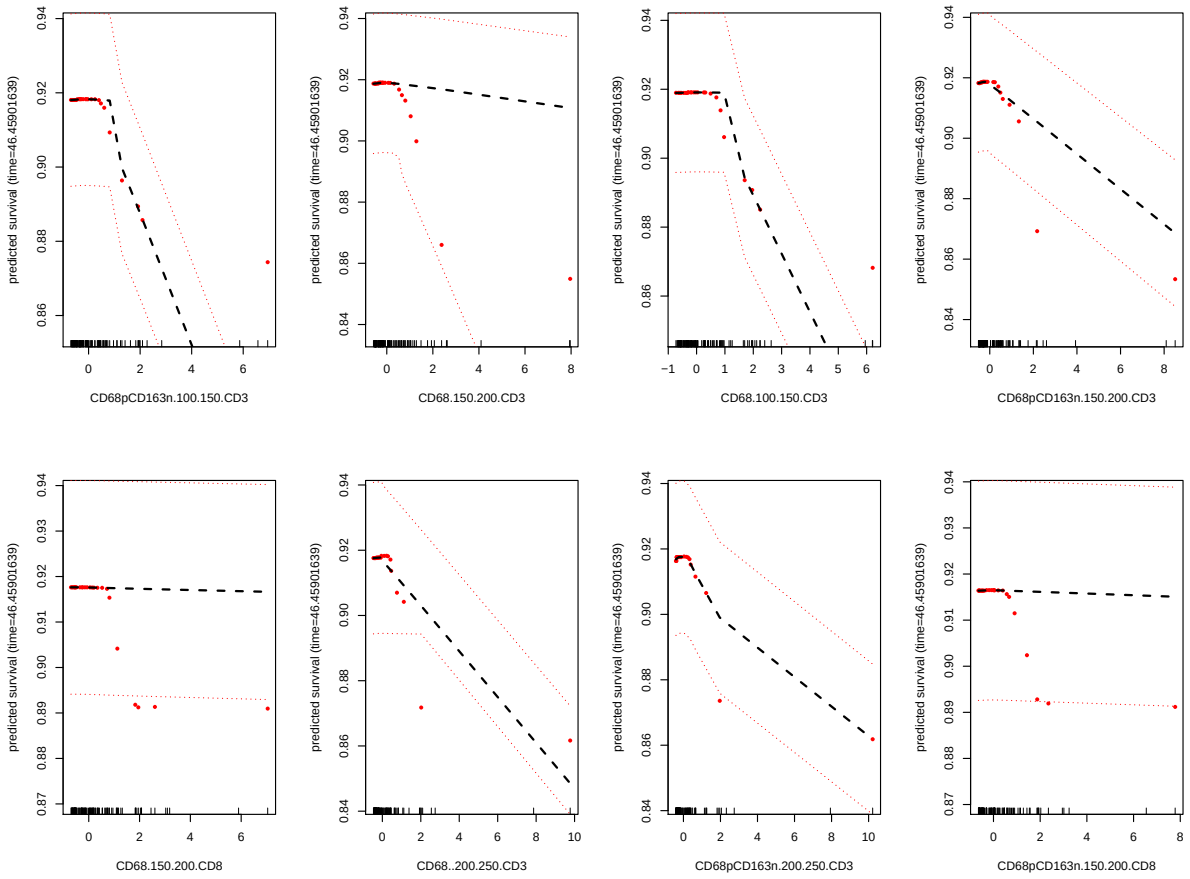


Figure 3.25: Partial plots for predicted survival at median follow up time ($t_{median} = 46.45901639$) (y-axis) versus top 8 variables (x-axis) in RF model trained with 5,000 trees, block size of 10 and terminal node size of 5. The dashed line denote the fitted RF model with 95% CI (red dashed boundary) and red dots as observed deaths. The red dots do not seem to be linear and not within the CI in several occasions.

According to figure 3.24, an arbitrary cutoff can be drawn after the top eight variables on account of the drop in variable importance. These variables were selected for partial dependence study on mortality. Most of the observations fall within the 95% CI range, as indicated by the red dashed boundary lines. For instance, the plots on the top left perform well while those in the bottom left less so. One of the dead patients (the red dot at the far right on each plot) is very likely to be an outlier as it is no near from the others. However, it is not certain as there are only 15 deaths in this training dataset.

From these plots, the non-linearity on the relationship between the survival and these variables can be illustrated. This RF model can capture some of these non-linearity but not all of them. At the median follow-up time, the general trend is that the predicted survival decreases as the values of these variables rises. It is also worth taking note of this edge effect of the 95% that there are

little observations on large values of these variables. This may limit the strength of this trend.

3.2.2.2 Gradient Tree Boosting

The following combination of parameters setting of the gbm survival model was trained with five-fold CV in the `gbm` package in `R`:

- The numbers of tree built were 10,000;
- The shrinkage were 0.001;
- The interaction depths, where is number of splits to be performed each tree, were 1 and 3.

The gbm models with interaction depth of 1 and 3 are very similar in terms of OOB error rates. For simplicity, the model with interaction depth of 1 was chosen as the optimal and it turned out to have better prediction performance on unseen validation dataset as shown in the next section.

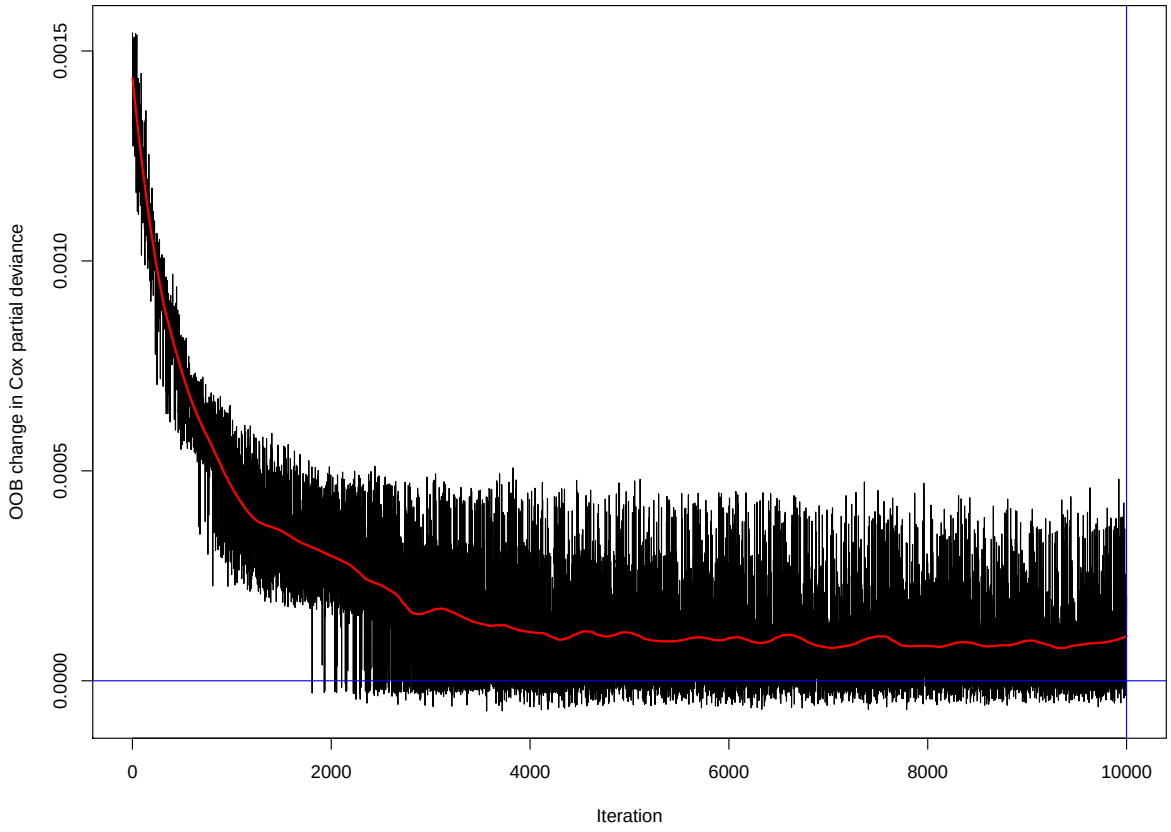


Figure 3.26: OOB partial deviance (y-axis) versus iteration (x-axis) plot for the gbm model trained with 10,000 trees, shrinkage of 0.001 and interaction depth of 1. The OOB error is measured in Cox partial deviance and iteration denotes the number of trees. The red line is the mean OOB error for the trees at certain iteration. The error rate first drops and then converges after 4000 iterations.

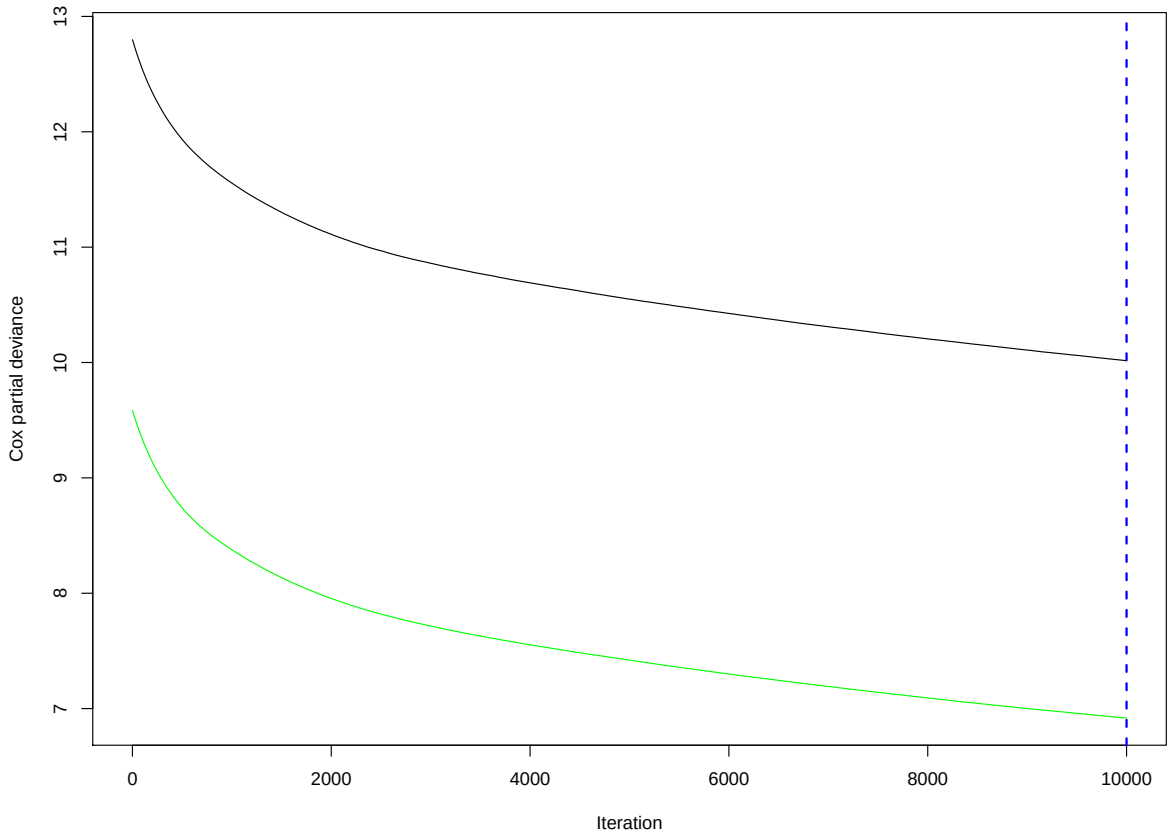


Figure 3.27: OOB error and CV in terms of Cox partial deviance (y-axis) versus iteration (x-axis) plot for the gbm model trained with 10,000 trees, shrinkage of 0.001 and interaction depth of 1. The green line denotes CV error and the black OOB. Both errors keep dropping until 10,000 iterations (blue dashed vertical line).

Figures 3.26 and 3.27 illustrates that OOB and five-fold CV deviance decrease as the iterations increase. It is apparent from this table that more iterations can attain a better CV score. However, considering the constraint of the computation run time and the fear of overfitting, it is reasonable to cap the number of trees up to 10,000 where the OOB error has already converged.

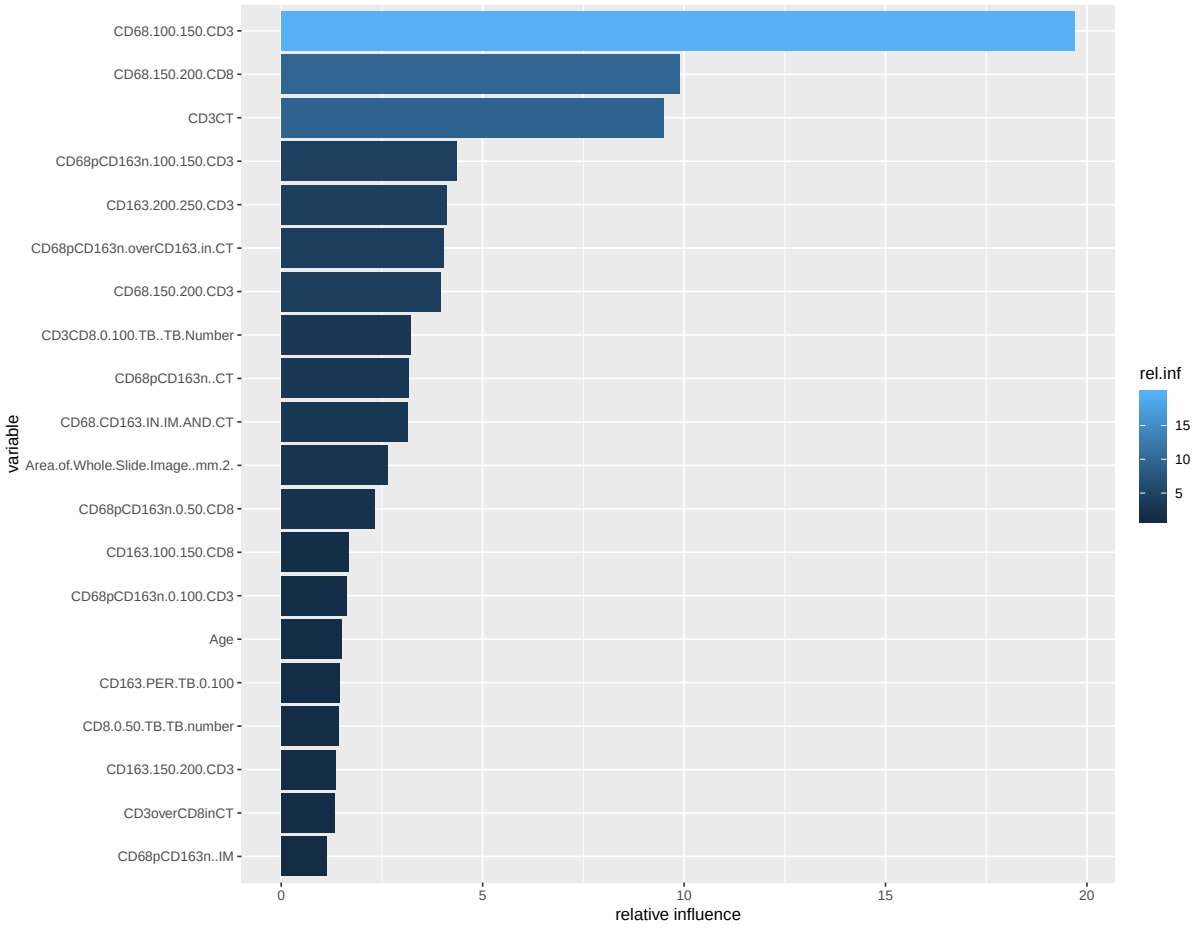


Figure 3.28: Variable importance (y-axis) versus top 20 predictors plot for gbm model trained with 10,000 trees, shrinkage of 0.001 and interaction depth of 1. The lighter the blue, the greater the relative influence of the variable (see legend). The top 3 predictors dominate most of the contribution to the gbm model.

The variable relative influence sorted in descending order are compared in figure 3.28. The significant variables, i.e. those having more than one in terms of relative influence, and corresponding values generated can be found in list A.9, Appendix A.

Similar to RF results, closer inspection of the table shows that the variables related to how macrophage infiltrates surround lymphocytic infiltrates from $100-200\mu m$ overall ranked high. These variables are two out of the top three. Among these, the importance of the variable describing the CD68 density in vicinity of CD8 within $100-150\mu m$ substantially stands out from the rest. Interestingly, this notable variable explains the density of CD68 but not at close range from CD3. This result was somehow counter-intuitive and may be worth further biological investigation. There is only one categorical variable listed, age group; age group ranks considerably lower than the continuous variables.

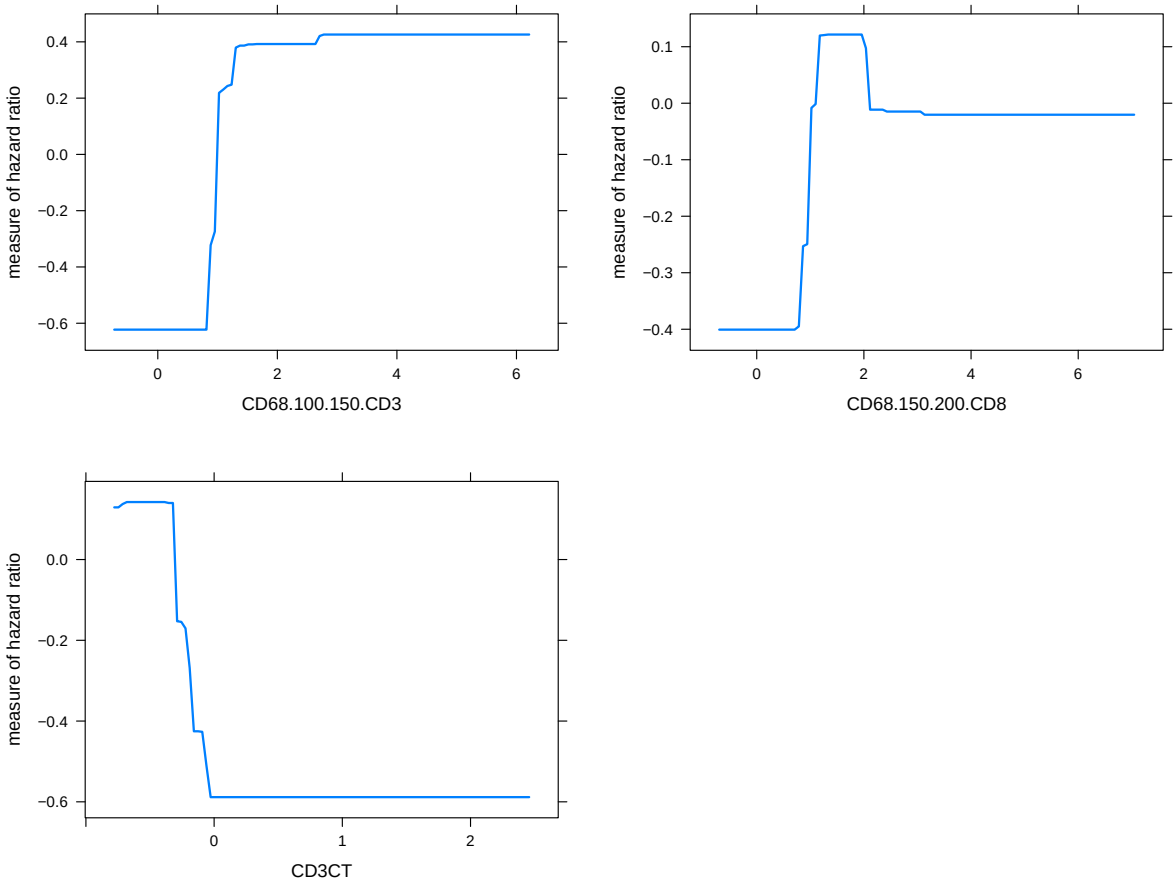


Figure 3.29: Marginal effect plot of Hazard Ratio (HR) versus the top 3 variables (x-axis) in gbm model trained with 10,000 trees, shrinkage of 0.001 and interaction depth of 1. The two on the top suggest that HR generally increases with these two predictors; HR decreases as CD3 in CT rises. All three lines do not look linear, in particular on the top right panel.

Figure 3.29 set out the marginal effects of the top three variables (CD68 within $100-150\mu m$ of CD3, CD68 within $100-150\mu m$ of CD8 and CD3 in CT) by "integrating" out the effects of other variables. These plots illustrate the non-linearity relationship. When CD68 near CD3 and CD8 increases, the risk of death increases generally. When CD3 in CT rises, the risk of death drops.

3.2.2.3 Predictions

	Validation(EDI)	Validation(JAP)
C-index	0.6242300	0.6526316
D_{xy}	0.2484600	0.3052632
S.D.	0.1326097	0.1716402
n	56	61
Missing	0	0
Uncensored	11	13
Relevant Pairs	974	1,140
Concordant	608	744
Uncertain	2,106	2,520

Table 3.7: Validation results for the optimal RF model

	Validation(EDI)	Validation(JAP)
C-index	0.6837782	0.7052632
D_{xy}	0.3675565	0.4105263
S.D.	0.1783700	0.1452244
n	56	61
Missing	0	0
Uncensored	11	13
Relevant Pairs	974	1,140
Concordant	666	804
Uncertain	2,106	2,520

Table 3.8: Validation results for the optimal gbm model

With reference to the C-index, the validation results of the optimal RF and gbm models in table 3.7 generally outperform those of regularised models in section 2.2. Although these are not considered to predict the cohorts very well (C-index < 0.8), these machine learning models can generalise to the validation cohorts that these C-indices were larger than 0.6. This may ascertain the non-linearity of the relationship.

In particular, given the constraint of lack of deaths in the dataset, the C-indices of this gbm survival model are close to 0.7, which indicate a fairly good model. In terms of C-index, it can be seen that the gbm model generally outperforms the rest in table 3.8. Similar to the RF

model, it serves slightly better for the Japanese validation cohort than the Edinburgh, which is unexpected since the training dataset comes from Edinburgh. It implies that the nature of the Japanese validation cohort is likely to resemble the training data. In this regard, it suggests that the machine learning models are not overtrained while having a satisfactory predictability.

Consolidating from all the results, there should be a subtle difference between these two validation cohorts. This may worth unsupervised learning or biological investigation.

It implies that the nature of the Edinburgh validation cohort is likely to resemble the train data, which was also obtained from Edinburgh. In this regard, these machine learning models are considered to be slightly overtrained.

3.3 Classification

In this section, the survival data is reduced to a classification problem, in order to sub-categorise patients into high and low risk groups. In other words, this analysis targets the classification of Disease Specific Death (DSD) while excluding DSS. The high risk group is the predicted death while the low risk group is the predicted censoring. Machines learning techniques were implemented with SMOTE oversampling and five-fold CV repeatedly for ten times, via the `caret` package in `R`. These machine learning models include NB, SVM and gbm. RF model was also implemented but the results were unsatisfactory. Therefore, it is not presented in this study.

To avoid overfitting, NB and SVM underwent variable selection based on the variable importance. All variables were fitted to obtain their variable importance. A forward selection process was then carried out by adding these variables one by one according to their importance, until the accuracy on the training dataset did not improve.

These models can predict censored (low risk) and deaths (high risk) on validation cohorts. The predictability of these models was assessed on the basis of confusion matrices, non-parametric KM plots and log-rank tests.

In order to take these rules further, a SMOTE oversampling on all three cohorts was undergone and then was split into a 70:30 training and validation data. This can project the classification power if more patients can be recruited.

3.3.1 Naïve Bayes (NB)

By forward selection, there is only one selected feature in NB classification model, which is CD68 within $150-200\mu m$ of CD3. For list of variable importance, see list A.10, Appendix A.

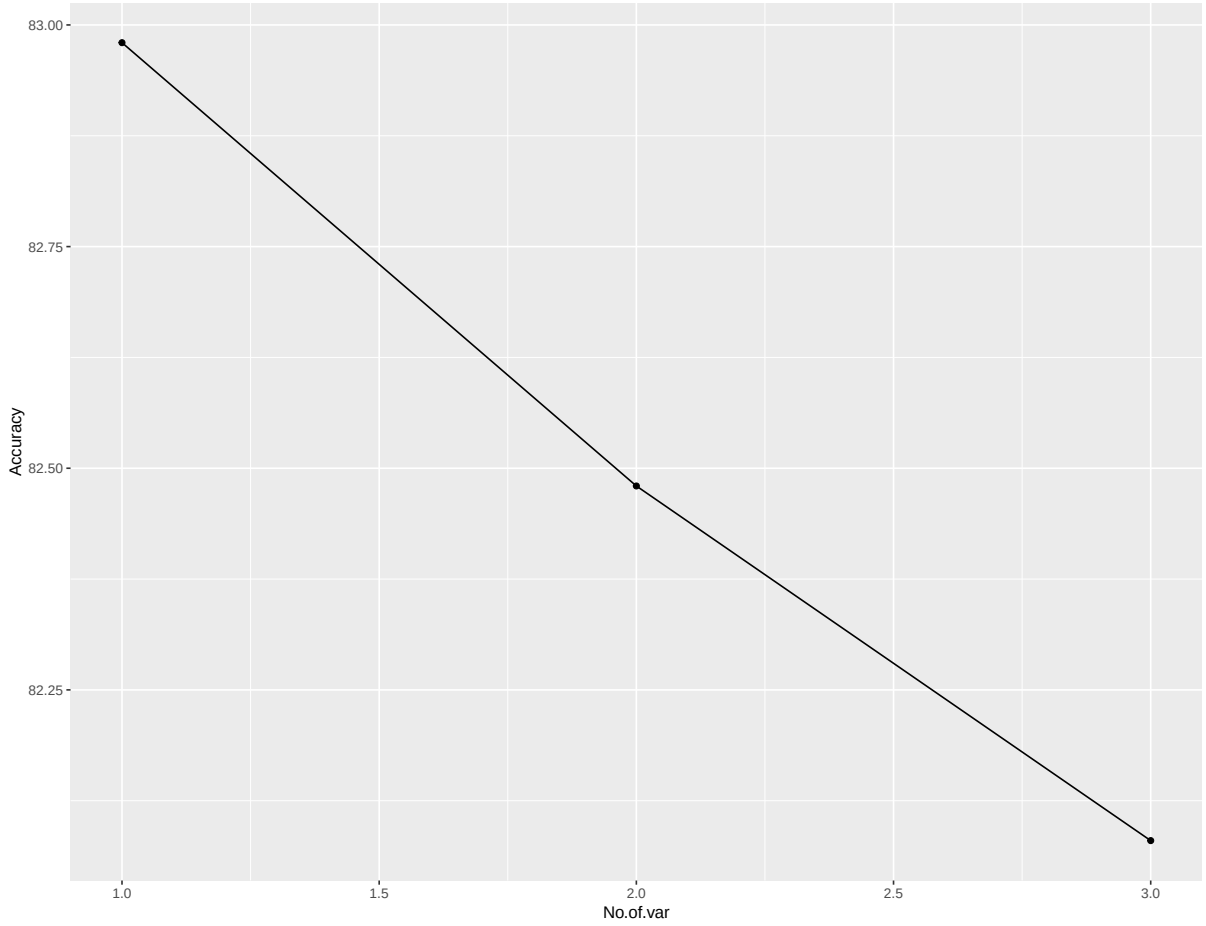


Figure 3.30: NB model accuracy (y-axis) versus number of fitted variables (x-axis) by forward selection. The accuracy starts to drop after fitting the first variable.

Despite only having one predictor, it performs surprisingly well with an accuracy of 82.98% on the training dataset classification. This supports NB as a simple but strong classifier. Thus, it can serve as a benchmark model for model comparison.

3.3.2 Support Vector Machines (SVM)

In view of the non-linearity observed in the survival models, radial basis kernel was used in constructing the SVM model. By forward selection, there are nine selected features (see figure 3.31 and table 3.9). The accuracy first rises quickly and then levels after the ninth variables.

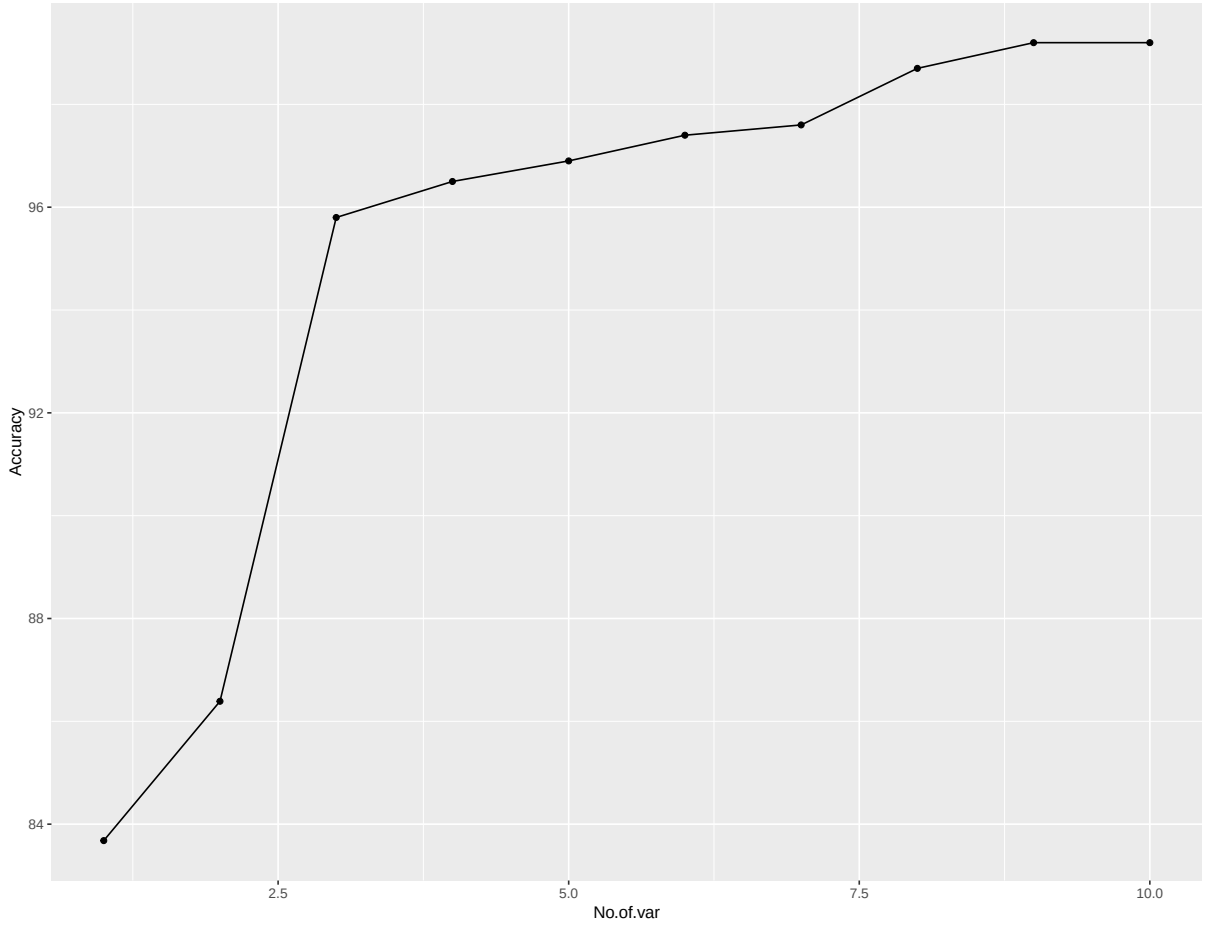


Figure 3.31: SVM model accuracy (y-axis) versus number of fitted variables (x-axis) by forward selection. The accuracy reaches maximum at the ninth variable with an accuracy of 99.2%.

It is worth noting that these features again are associated with the linkage between macrophage and lymphocytic infiltrates. This again merits further biological investigations on how this proximity information influences the CRC survival. Although most of the features are highly correlated, SVM is able to deal with their interactions while not affecting the accuracy.

SVM feature selection	
CD68 within $150-200\mu m$ of CD3	CD68 within $200-250\mu m$ of CD3
$CD68^{+}CD163^{-}$within $150-200\mu m$ of CD3	CD68 within $100-150\mu m$ of CD3
CD3 in CT	Ratio of $CD68^{+}CD163^{-}$ $200-250\mu m$ of CD3
$CD68^{+}CD163^{-}$ within $100-150\mu m$ of CD3	CD68 within $150-200\mu m$ of CD8
$CD68^{+}CD163^{-}$ within $150-200\mu m$ of CD8	
Ranked in descending order according to variable importance	(from top to bottom, left to right) [2]

Table 3.9: Parameters which appeared in the SVM model with budget (**C** = 32)

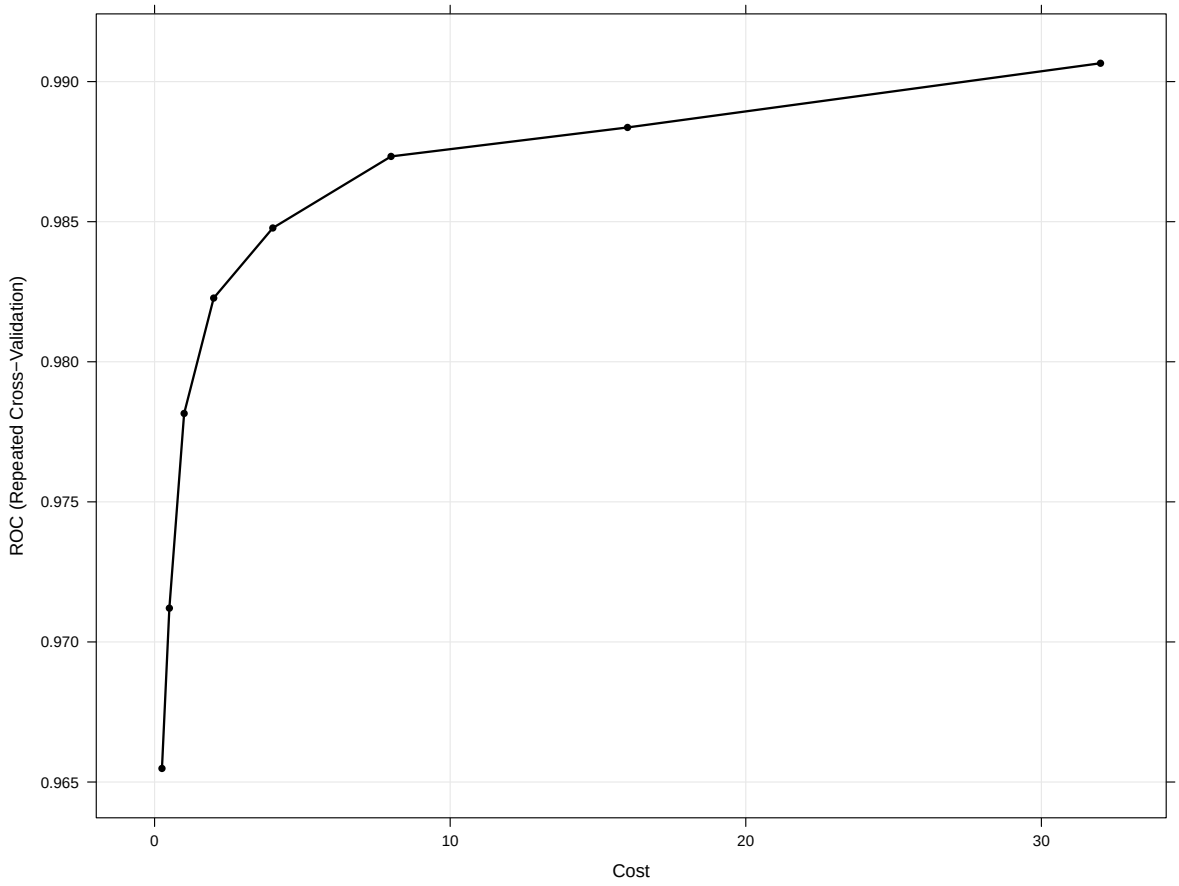


Figure 3.32: Hyperparameter tuning (Cost in x-axis) for 5-fold CV in terms of AUROC (y-axis) of SVM models using radial basis kernel. AUROC can formulate the capacity of how well a model can classify. The AUROC increases with cost and reaches maximum at cost=32.

Figure 3.32 illustrates how the hyperparameter, i.e. budget in SVM, was tuned via the AUROC metric. The optimal model was selected in accordance to the best AUROC during CV (see A.11, Appendix A). The cost was tuned to 32 for the SVM model with a striking accuracy of 99.2%. The high cost indicates high tolerance and this implies complex overlapping and interactions of the data points.

For illustration, figure 3.33 shows how SVM copes with the complexity of the data point distribution. This visualisation was compiled via the `e1071` package in `R` [35]. These variables are among the nine selected features. Though these variables should be correlated as they are both linked to CD3, the radial basis kernel is able to provide extra flexibility, i.e. support vector denoted by "x", to capture the signals.

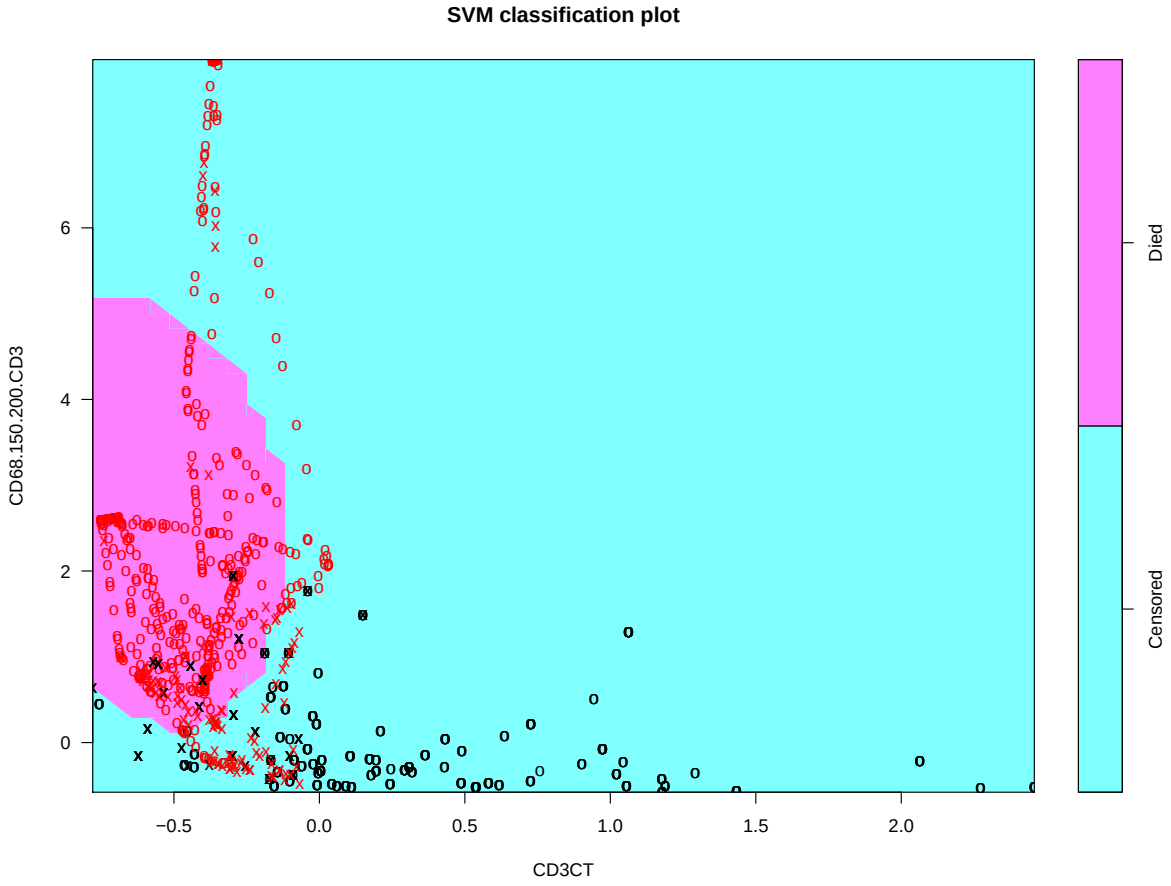


Figure 3.33: 2D partition plots among the interactions of 2 selected features (x- and y-axis) in SVM, i.e. CD68 within $150-200\mu m$ of CD3 (y-axis) versus CD3 in CT (x-axis). The "x" denotes the points that affect the support vectors with cost=32 while "o" is other normal observation; the black denote censored and red for death. The decision boundary of SVM is shown by the pink and blue areas. SVM performs well to delineate the interaction between these significant predictors. This is only

3.3.3 Gradient Boosting

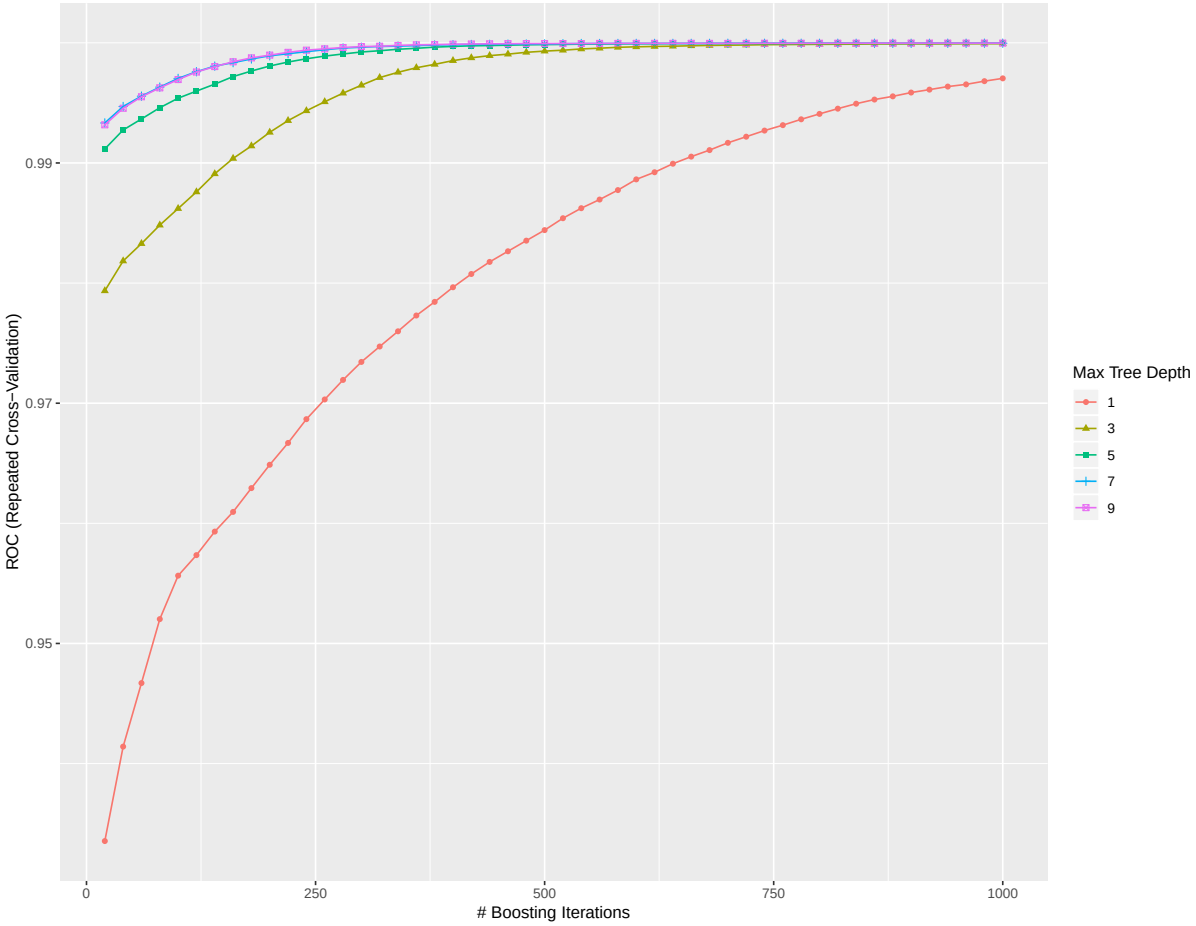


Figure 3.34: Hyperparameter tuning for 5-fold CV in terms of AUROC (y-axis) versus iteration (x-axis) of SVM models using radial basis kernel.

The following combination of parameters setting of the gbm classifications model was trained with five-fold CV:

- The numbers of tree, i.e. iteration, built were every 20 trees up to 1000;
- The shrinkage were 0.01;
- The interaction depths, where is number of splits to be performed each tree, were 1, 3, 5, 7 and 9;
- The minimum number of observations in terminal nodes is 20.

AUROC for maximum tree depth of 5, 7 and 9 converges after 500 iterations. This indicates that these gbm have become stable. The hyperparameter tuning for maximum tree depth and the

number of trees is presented in figure 3.34. The final gbm model with the best AUROC has 880 trees and interaction depth of 9. Interaction depth of 9 indicates the complexity of the features due to collinearity. Nonetheless, the performance of gbm is outstanding that it can correctly classify all observations in this training dataset. However, this raises concerns on overtraining the dataset.

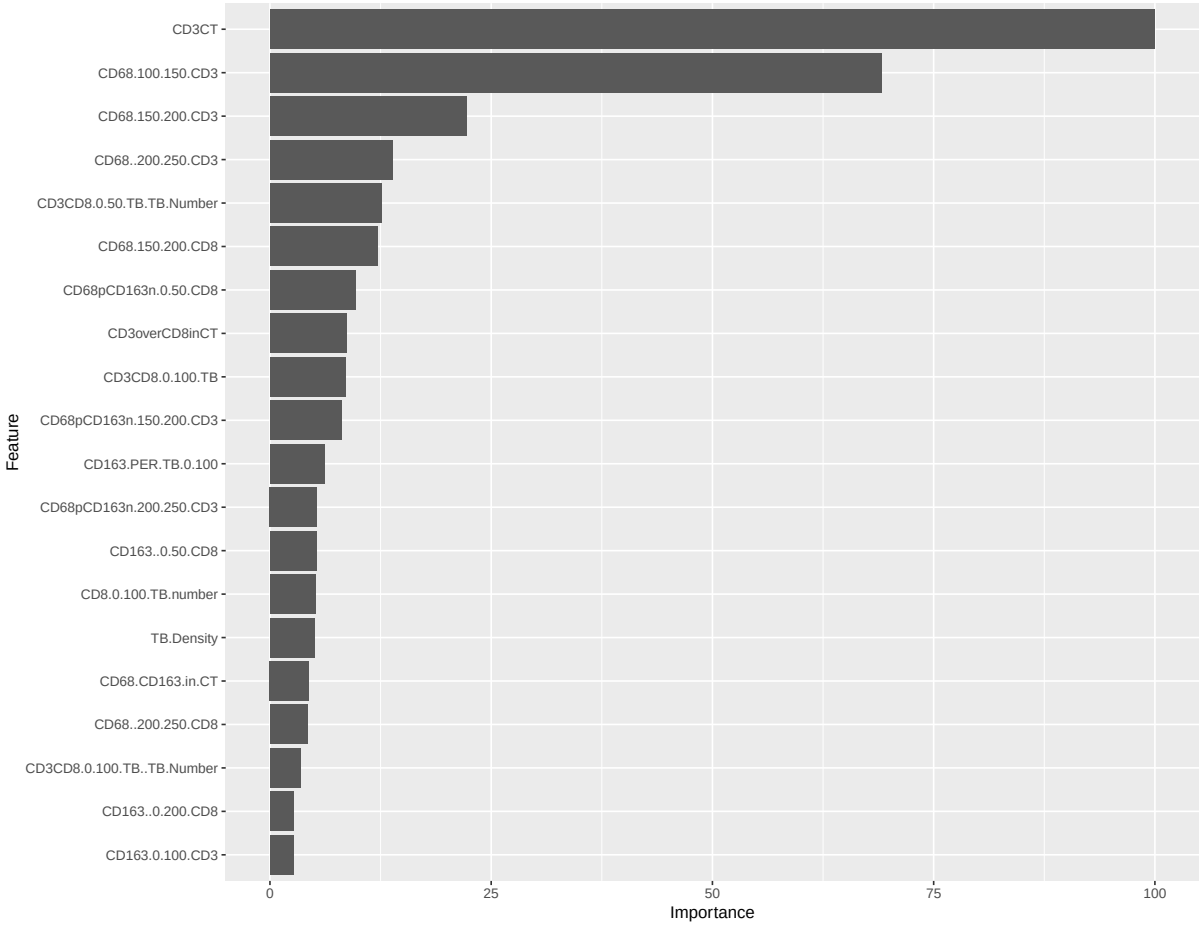


Figure 3.35: Variable importance (x-axis) plot for top 20 variables (y-axis) in gbm model trained with 880 trees and interaction depth of 9. The top 2 variables dominate the contribution to gbm.

Interestingly, the variable importance of this gbm model, as shown in figure 3.28, differs from that for survival analysis, as well as NB and SVM classification models. Though there is still a large component of macrophage infiltrate among the top twenty important features, this gbm model puts more emphasis on the lymphocytic infiltrates, especially CD3. For example, the most important variable is CD3 in CT. Together with the second most important variable, CD68 within $100-150\mu m$ of CD3, these variables appear to outweigh the rest.

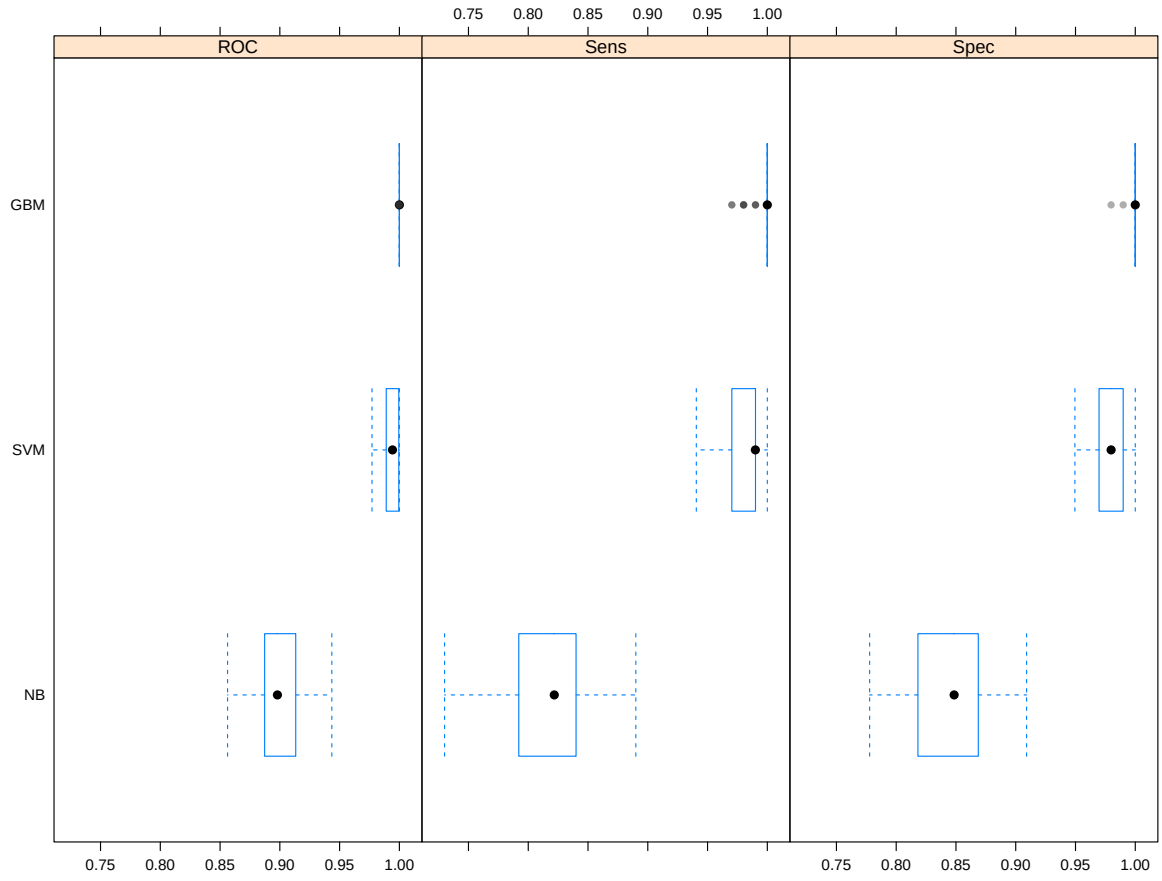


Figure 3.36: AUROC, sensitivity and specificity comparison (x-axis) across NB, SVM and gbm models (y-axis). The boxplots indicate their performances via resampling distribution [36]. gbm is the best, seconded by SVM and finally NB.

A model comparison is generated in figure 3.36 via their resampling distributions [36]. The gbm model outperforms the NB and SVM models, in terms of AUROC, sensitivity and specificity. Its 95% CI is very narrow in these resampling distributions. Nonetheless, there is concerns on overfitting of this gbm model. On the other hand, the performance of SVM is also satisfactory, though the specificity performs slightly better than the sensitivity. As a benchmark model, NB model has lower AUROC, sensitivity and specificity, as well as a much wider 95% CI than the other two models.

3.3.4 Predictions

In this section, the prediction performance of those models tested on the Edinburgh and Japanese cohorts are presented. The accuracy, sensitivity, specificity,KM plot and log-rank test based on these classification rules are compared across models and the current golden prognosis standard,

pT stage. Assessment on predictability based on accuracy, sensitivity and specificity can be tricky in this study. On one hand, those censored subjects are not exactly equivalent to those who did not die. On the other hand, owing to the imbalance, failing to correctly predict one dead patient would cost a great loss in sensitivity. Nonetheless, these measures can provide objective references to the model performance. Projections of these models with a SMOTE oversampling are presented at the end of this section.

3.3.5 Accuracy, sensitivity and specificity

Since medical data is often imbalanced that most of the subjects are censored, accuracy cannot be the one and only measure to assess the predictability. By the same token, fitted models can have capacity to obtain better specificity than sensitivity. These results are generated from the confusion matrices (see tables B.4 to B.11, Appendix B).

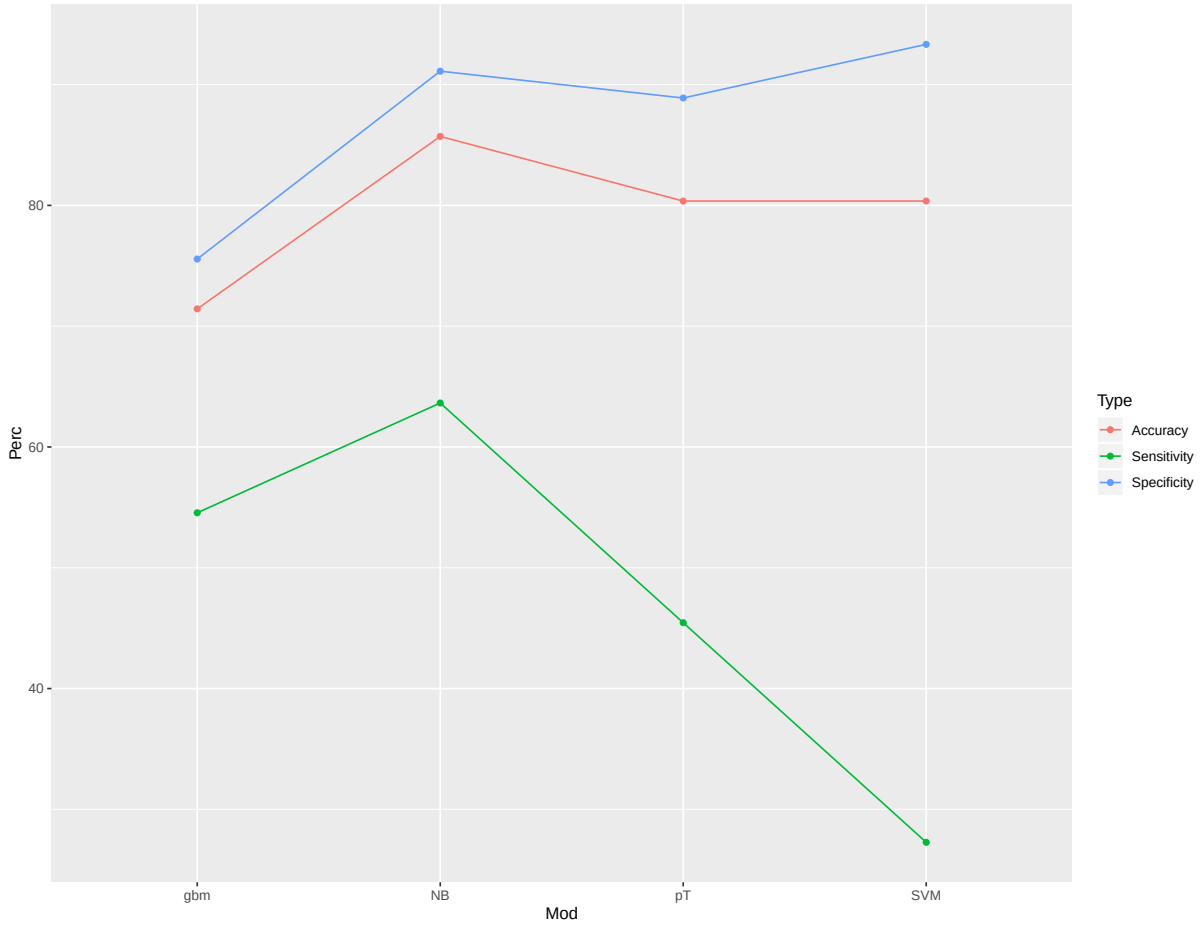


Figure 3.37: Accuracy, sensitivity and specificity (y-axis) across NB, SVM, gbm and pT Stage (x-axis) on Edinburgh validation cohort. NB performs remarkably well here with the best sensitivity of 63.64% while keeping a fairly high accuracy.

Regarding the validation results of the Edinburgh and Japanese cohort, all four classification rules do not work very well for sensitivity as shown in figures 3.37 and 3.37.

For the Edinburgh cohort, the NB model, which performs the best, has a sensitivity of 63.64%. It highlights the potential how much one predictor can explain with a simple NB model. SVM achieves a higher specificity at the expense of sensitivity. Although the sensitivity is not bad, gbm here does not perform as well as the others.

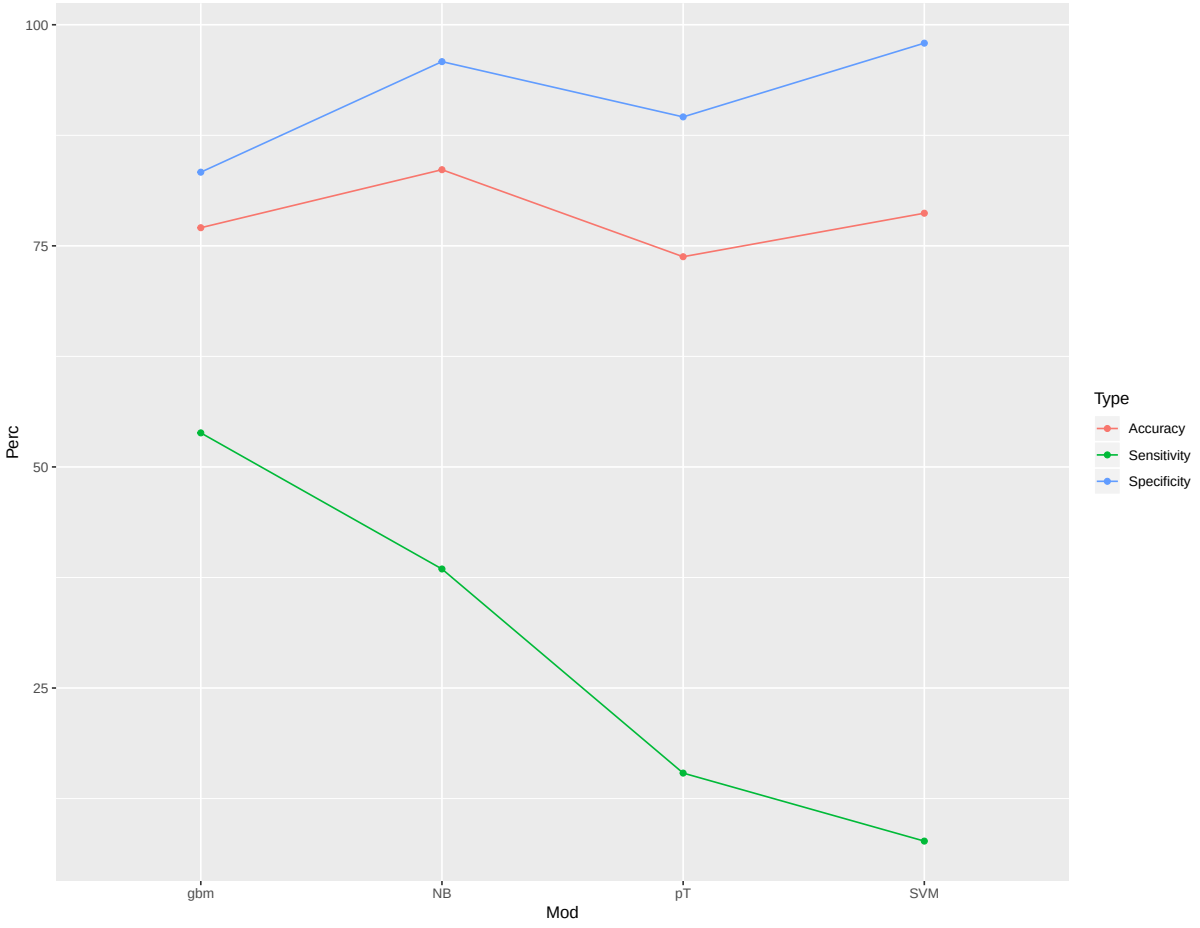


Figure 3.38: Accuracy, sensitivity and specificity (y-axis) across NB, SVM, gbm and pT Stage (x-axis) on Japanese validation cohort. Gbm performs the best here averaging all three metrics.

With respect to the validation results of the Japanese cohort, the gbm stands out for a sensitivity of 53.85% while having a second best accuracy of 77.05%. In spite of having the highest accuracy (83.61%), NB sacrifices sensitivity. SVM and pT Stage do not perform well here.

3.3.5.1 Kaplan-Meier (KM) plot and log-rank test

The KM plot is a non-parametric tool to assess the survival function from time and event data. This can visualise the proportion of alive patients with respect to follow-up time. For detailed console output in this section, see list A.13, Appendix A.

Figures 3.41 to 3.46 are the KM plots for each classification rule. For pT stage, it performs well in the Edinburgh validation cohort but not in the Japanese. Regarding NB, though it only classifies a small portion of subjects into the high risk group, the log-rank tests and KM plots agree the difference between the high and low risk groups. With respect to SVM, its performance on the Edinburgh validation cohort is significant but it does not replicate its success in the Japanese. For gbm, it performs very well in both cohorts, while its performance in the Edinburgh validation cohort is better than in the Japanese. None of these rules have a p-value smaller than 0.05 for the Japanese validation cohorts. Overall, gbm classification rule stands out from the rest.

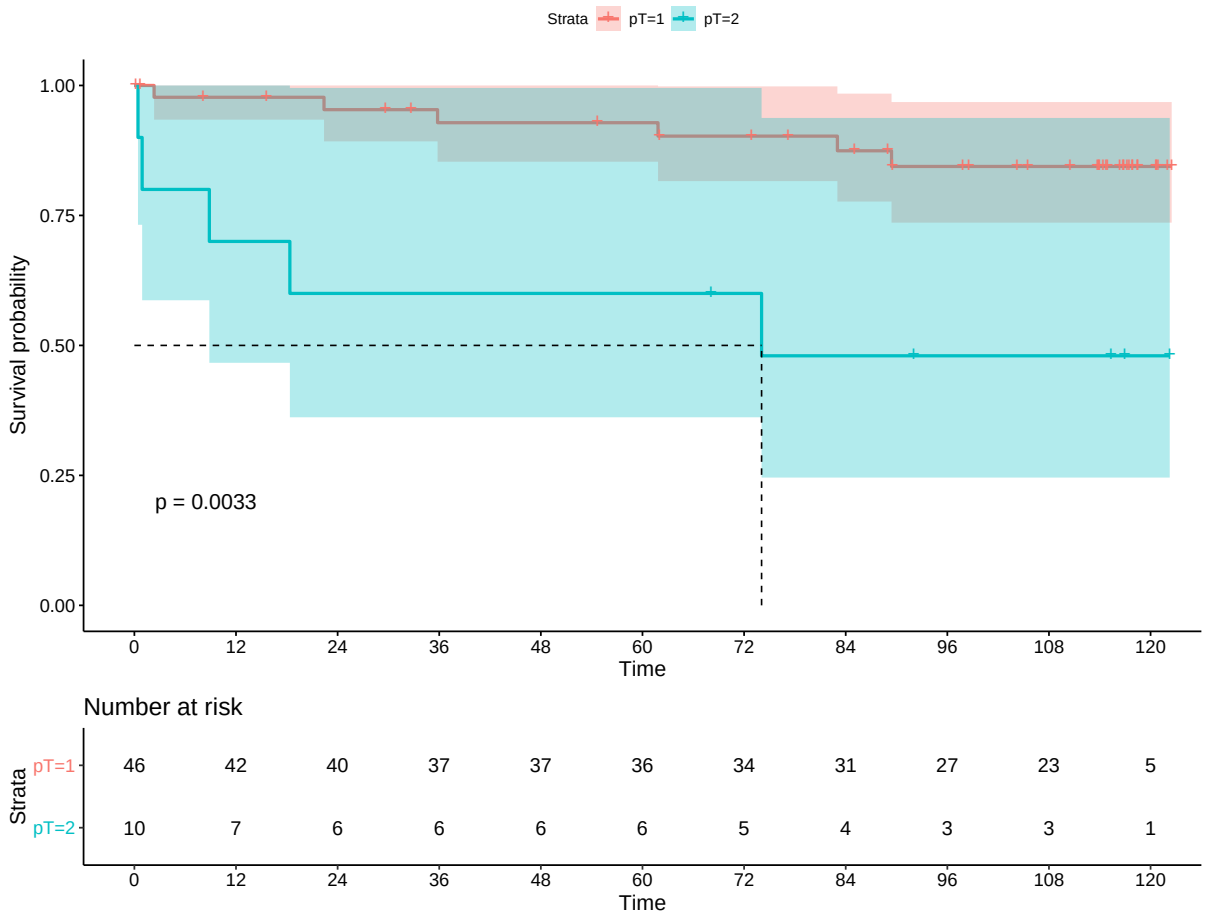


Figure 3.39: KM plot of survival probability (y-axis) versus time (x-axis) for pT Stage on Edinburgh validation. The censored patients are marked with a plus sign (+) on the curve. The dashed line denotes the median survival. The 95% CIs are printed on the plots. In lack of dead patients' information, the CIs for the high risk groups are expected to be wide. The p on the plot represents the p-value of log-rank test, which is a non-parametric hypothesis test for the group difference of survival sample distributions. A risk table is attached under each plot to illustrate the breakdown of censored and dead patients, in terms of patients at risk, at every twelve months. pT Stage can identify patients with lower survival chance with a p-value of 0.0033 in log-rank test, despite the large area of overlapping in terms of the CI. The pT3 group (pT=1) has survival chance above 0.85 for over 120 months.

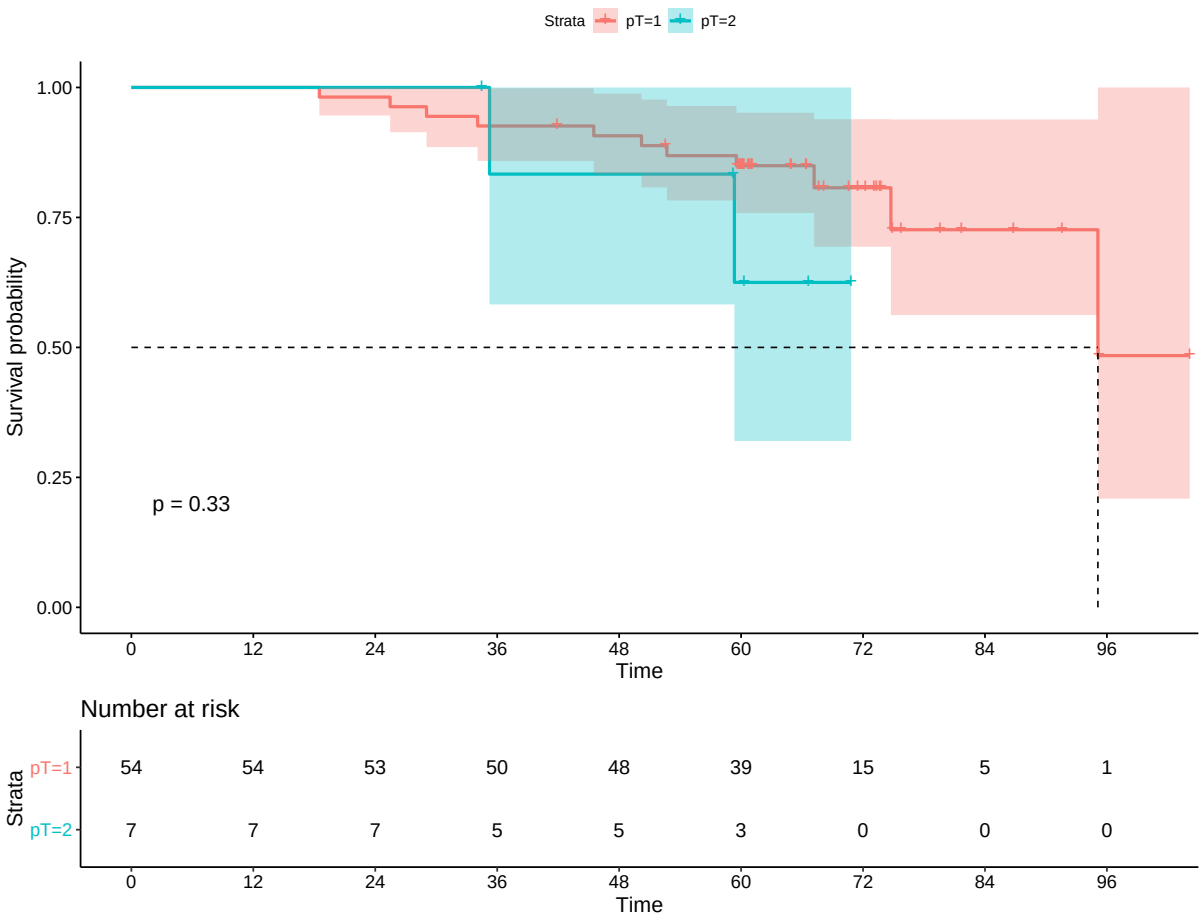


Figure 3.40: KM plot of survival probability (y-axis) versus time (x-axis) for pT Stage on Japanese validation. pT Stage cannot identify patients with lower survival chance with a p-value of 0.33 in log-rank test.

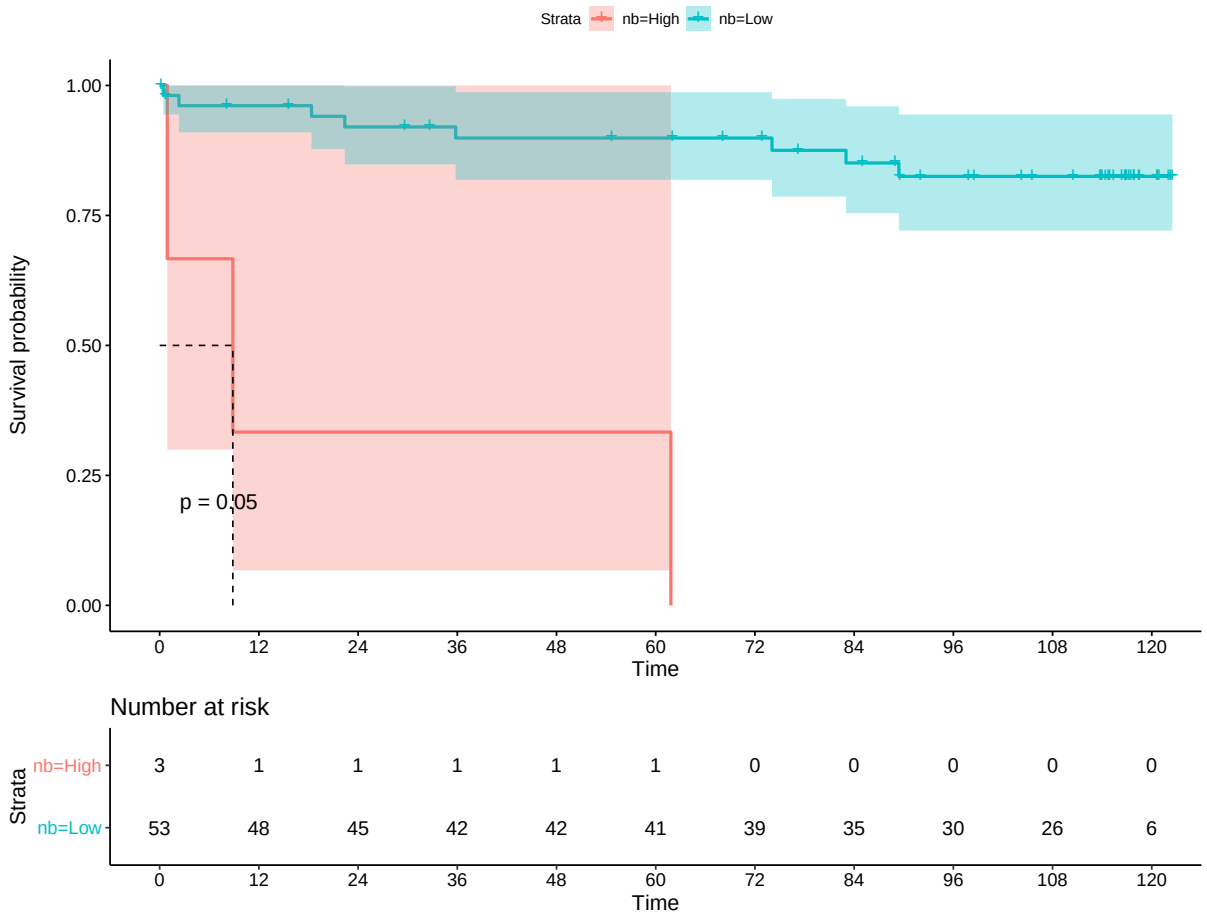


Figure 3.41: KM plot of survival probability (y-axis) versus time (x-axis) for NB on Edinburgh validation. NB can identify patients with lower survival chance with a median survival time of less than 12 months and a p-value of 0.05 in log-rank test, despite only 3 patients in the high risk group. The low risk group has survival chance above 0.8 for over 120 months.

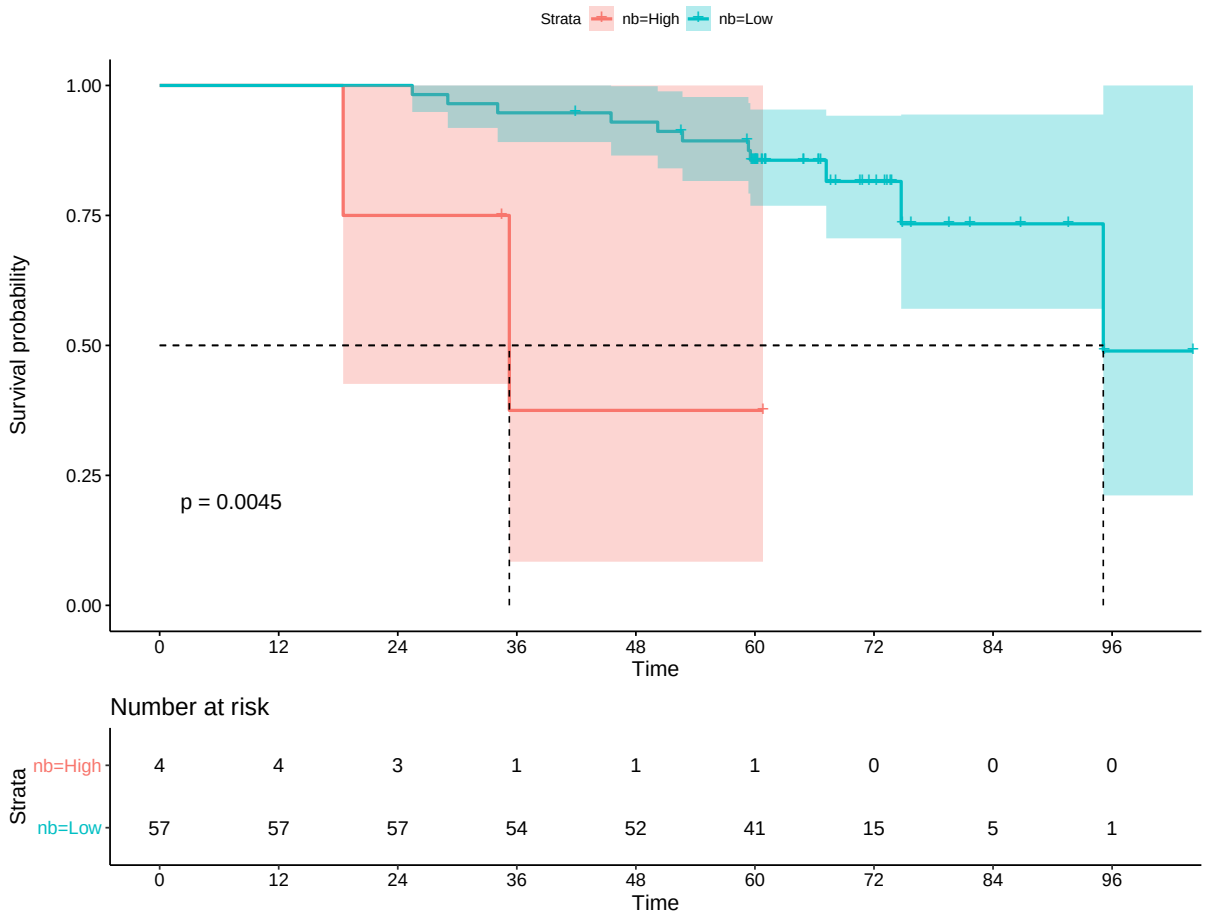


Figure 3.42: KM plot of survival probability (y-axis) versus time (x-axis) for NB on Japanese validation. NB here can identify patients with lower survival chance with median slightly less than 36 months with a p-value of 0.0045 in log-rank test, despite only 4 patients in the high risk group. The low risk group can have survival chance less than 0.5 at around 96 months.

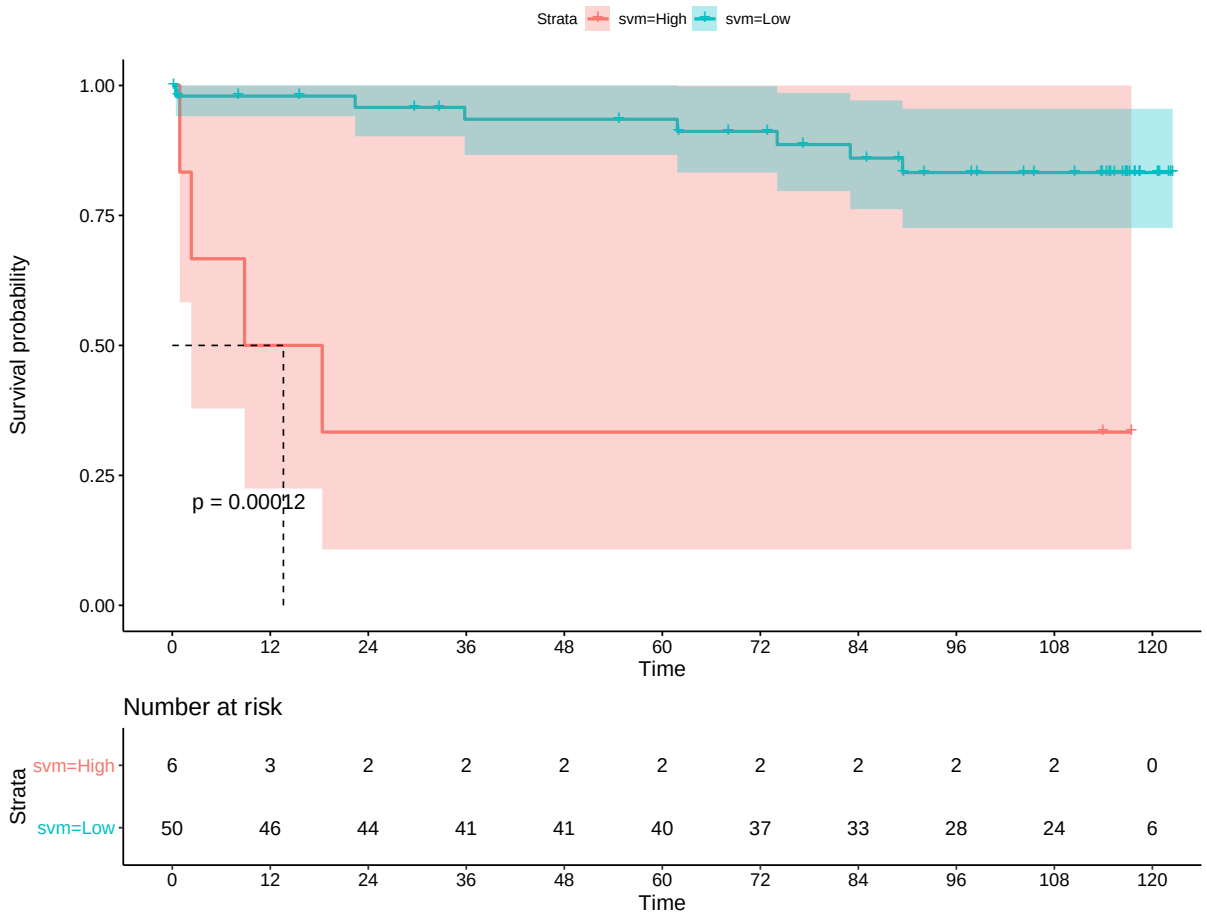


Figure 3.43: KM plot of survival probability (y-axis) versus time (x-axis) for SVM on Edinburgh validation. SVM here can identify patients with lower survival chance with median slightly more than 12 months with a p-value of 0.00012 in log-rank test while keeping the chance above 0.8 for low risk group.

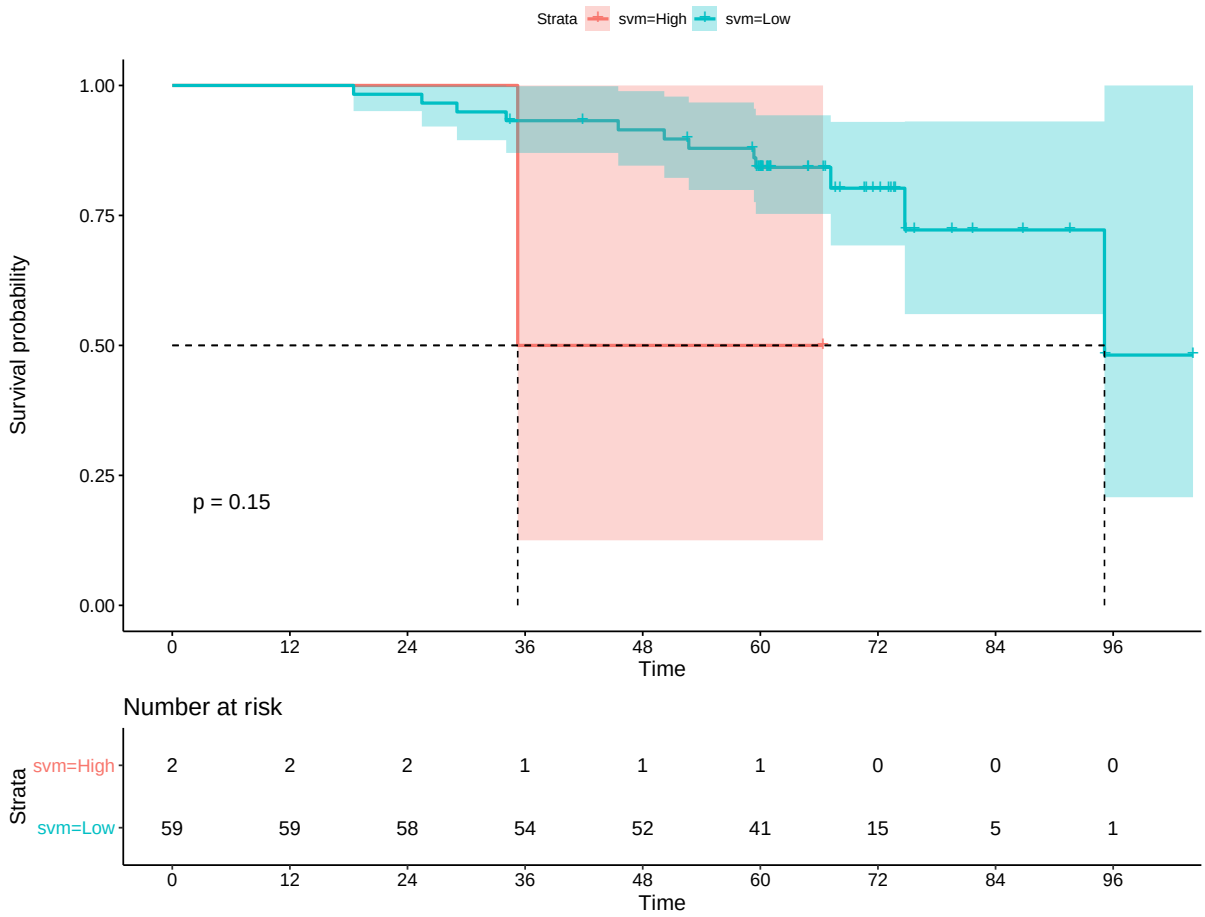


Figure 3.44: KM plot of survival probability (y-axis) versus time (x-axis) for SVM on Japanese validation. SVM here cannot identify patients with lower survival chance with a p-value of 0.15 in log-rank test and only two patients in the high risk group.

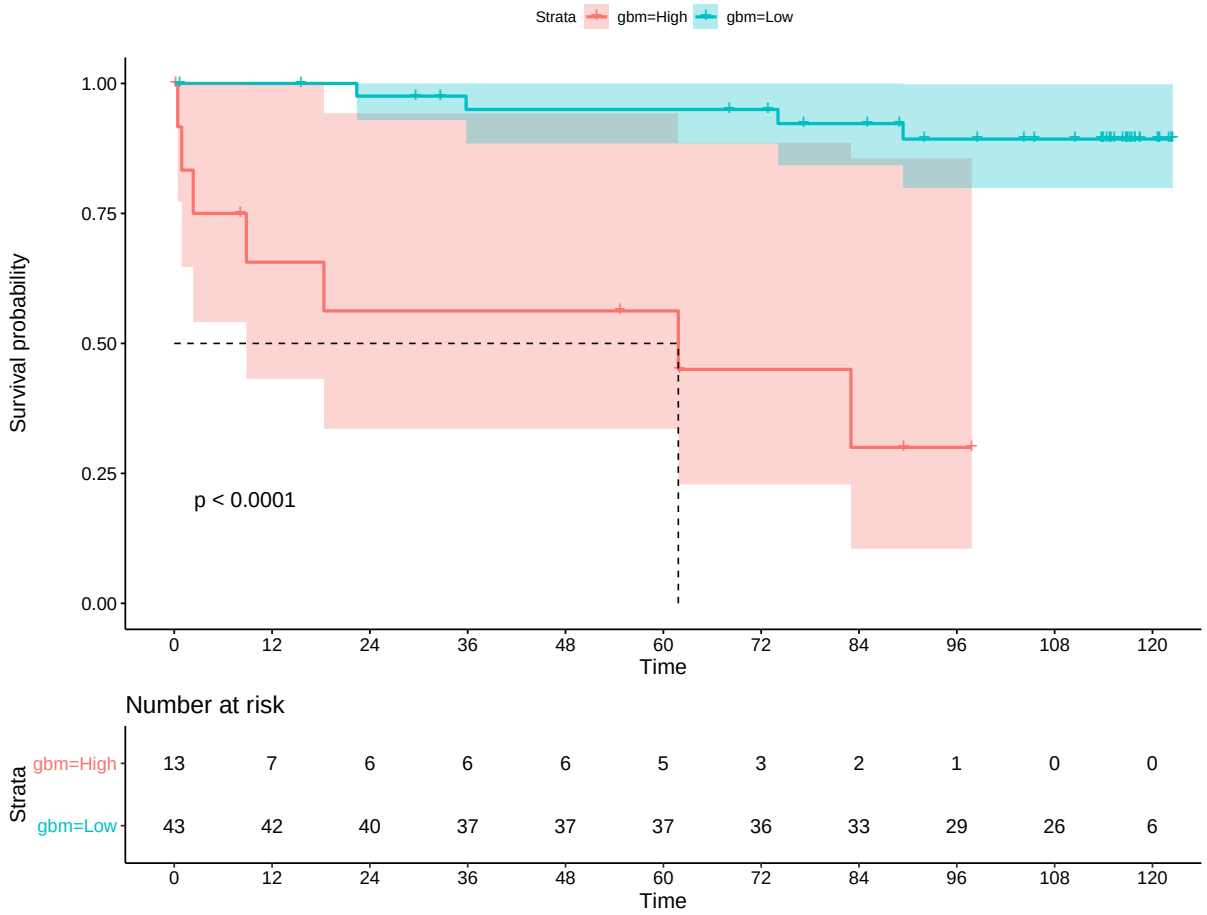


Figure 3.45: KM plot of survival probability (y-axis) versus time (x-axis) for gbm on Edinburgh validation. Although the log rank test has a p-value of 0.036, gbm does not perform well with respect to how close these groups are throughout the time. The low risk has a survival chance above 0.75 after 120 months.

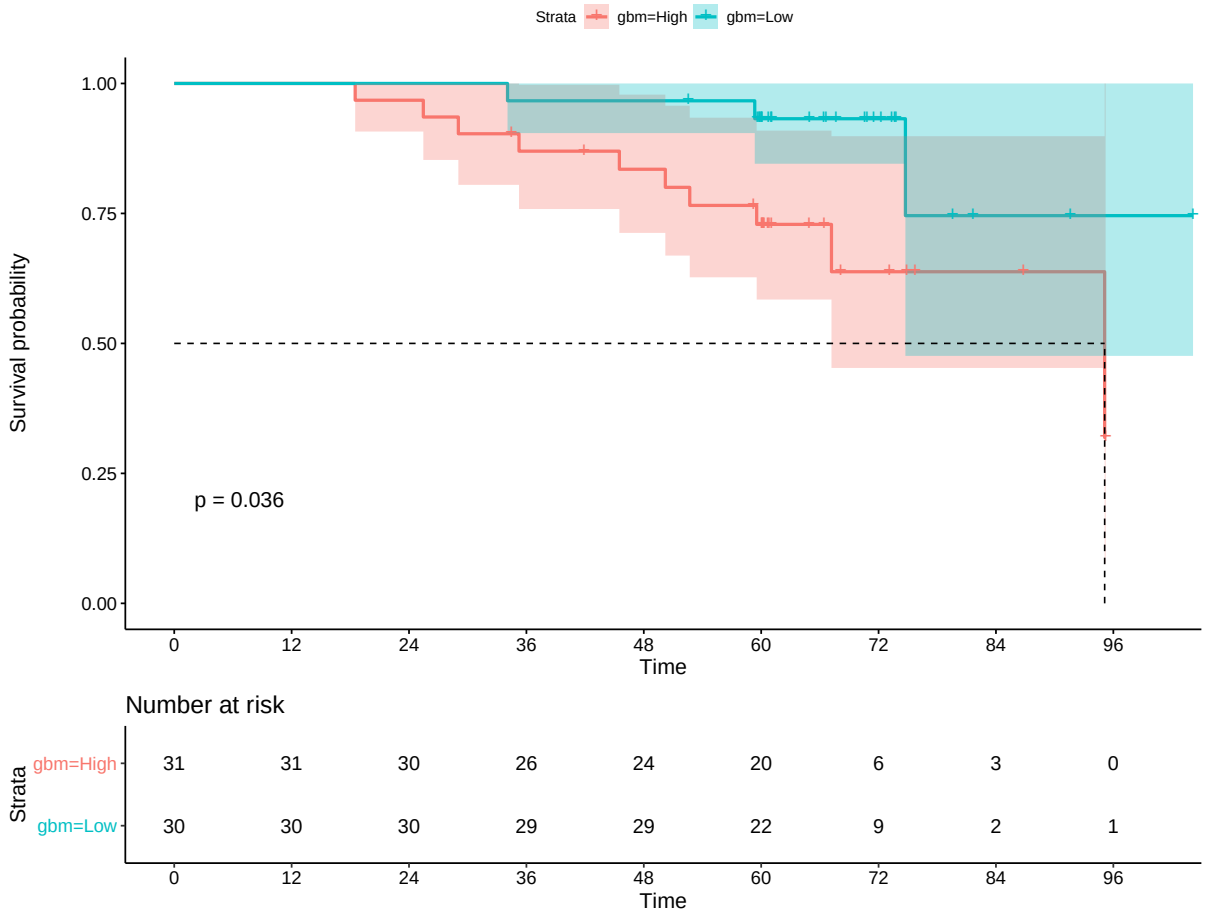


Figure 3.46: KM plot of survival probability (y-axis) versus time (x-axis) for gbm on Japanese validation. gbm here can identify patients with lower survival chance with median slightly more than 60 months with a p-value <0.0001 in log-rank test, while keeping the survival chances almost 0.9 for low risk group.

3.3.5.2 Projection

It is hard to judge if the models are well-performed in these cohorts since there is no consistency for any models to perform satisfactorily. It is suspected either there are not enough observations to train the model; or there is unknown difference between the Edinburgh and Japanese cohorts that causes the misclassification. Moreover, the three metrics can hugely vary due to the small cohort sizes, i.e. a few correctly predicted observations can make a huge difference in accuracy. Furthermore, owing to the class imbalance, accuracy favours specificity than sensitivity. Therefore, an oversampled dataset can offer a more objective perspective to assess the performances of these methods.

SMOTE oversampling takes all three cohorts and bootstrap the entire dataset to a sample size of $n = 986$, with $n_{censored} = 830$ and $n_{dead} = 156$ via k-NN=5. This oversampled dataset is

approximately proportional to the original dataset. This rest on the assumption that these three cohorts are representative samples from the entire population.

Figure 3.47 summarises the general assessment metrics for each rule. All classification rules have a p-value smaller than 0.05 for this oversampled validation dataset. gbm has excellent performance in both accuracy and specificity. Both SVM and gbm have outstanding performances according to the KM plots (see figures 3.50 and 3.51), with median survivals of approximately thirty-six months for the high risk group. In particular, SVM only uses nine highly correlated predictors. For pT group, there is some overlapping of the 95% CI between the pT3 and pT4 Stages. Comparing figures 3.48 and 3.49, NB performs better than the pT Stage classification with only 8 patients in high risk group.

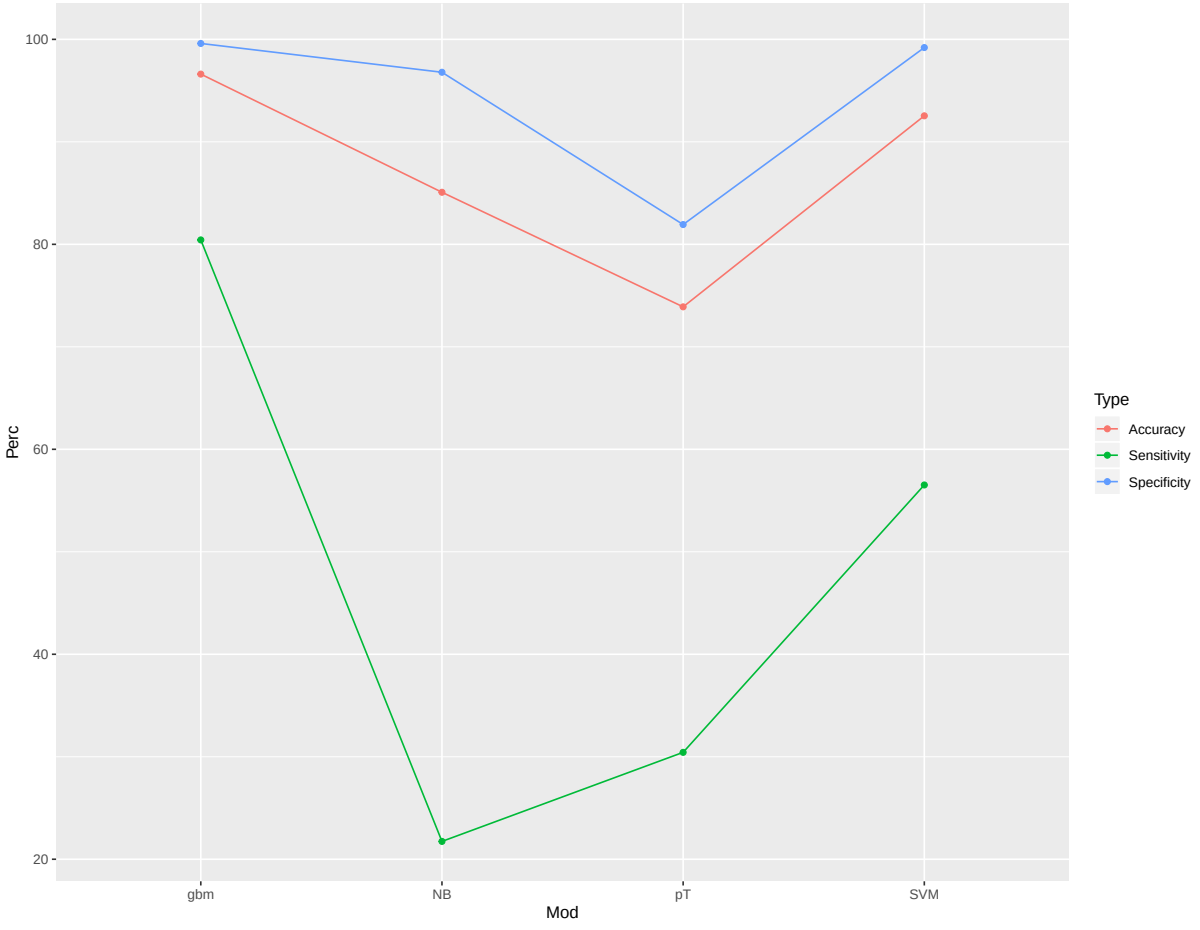


Figure 3.47: Accuracy, sensitivity and specificity (y-axis) across NB, SVM, gbm and pT Stage (x-axis) on oversampled validation dataset. Gbm, SVM both perform well while the former has a much higher sensitivity than the latter.

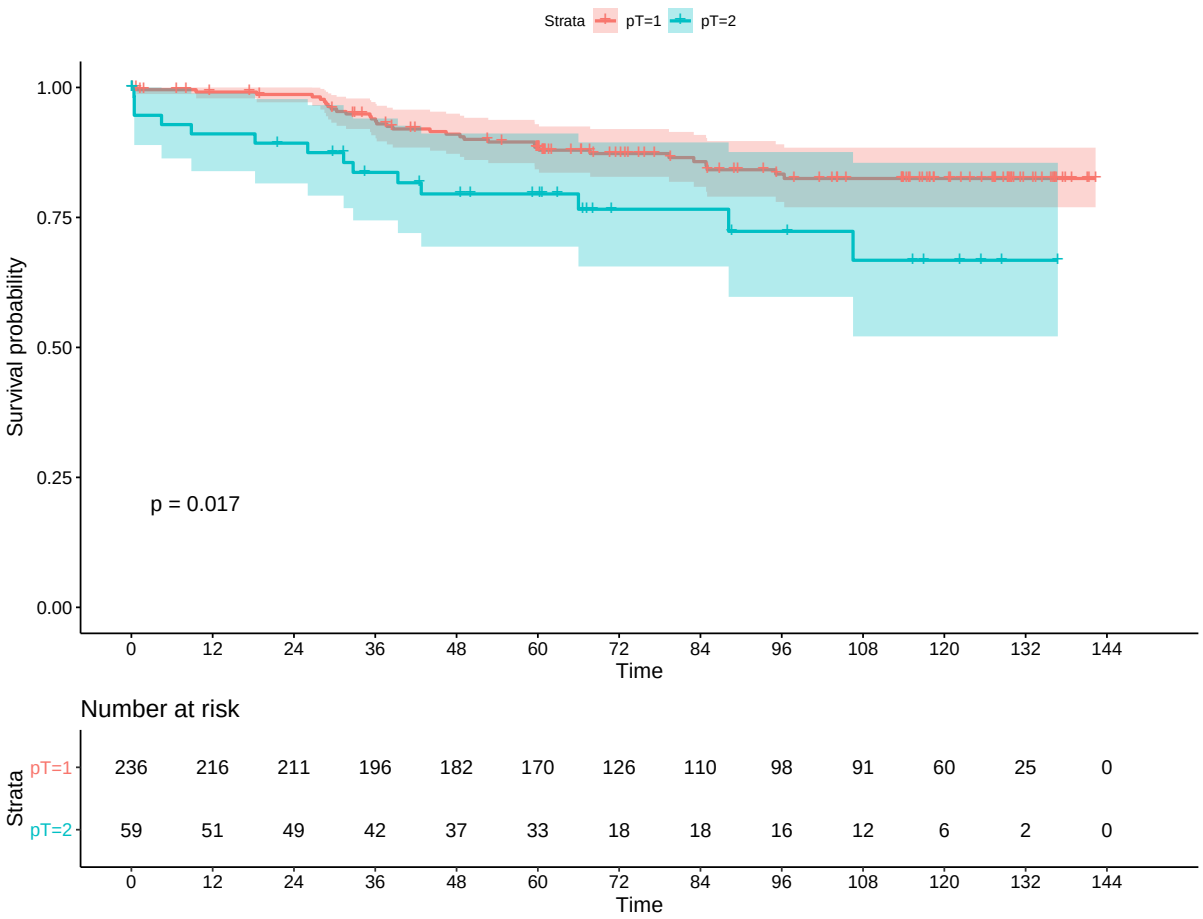


Figure 3.48: KM plot of survival probability (y-axis) versus time (x-axis) for pT Stage with oversampling. Despite having a p-value of 0.017 in log rank test, pT Stage does not perform well as CIs of these groups overlap most of the time.

		Predicted	
		pT3	pT4
Actual	Censored	204	45
	Dead	32	14

Table 3.10: Confusion matrix for pT Stage with oversampling

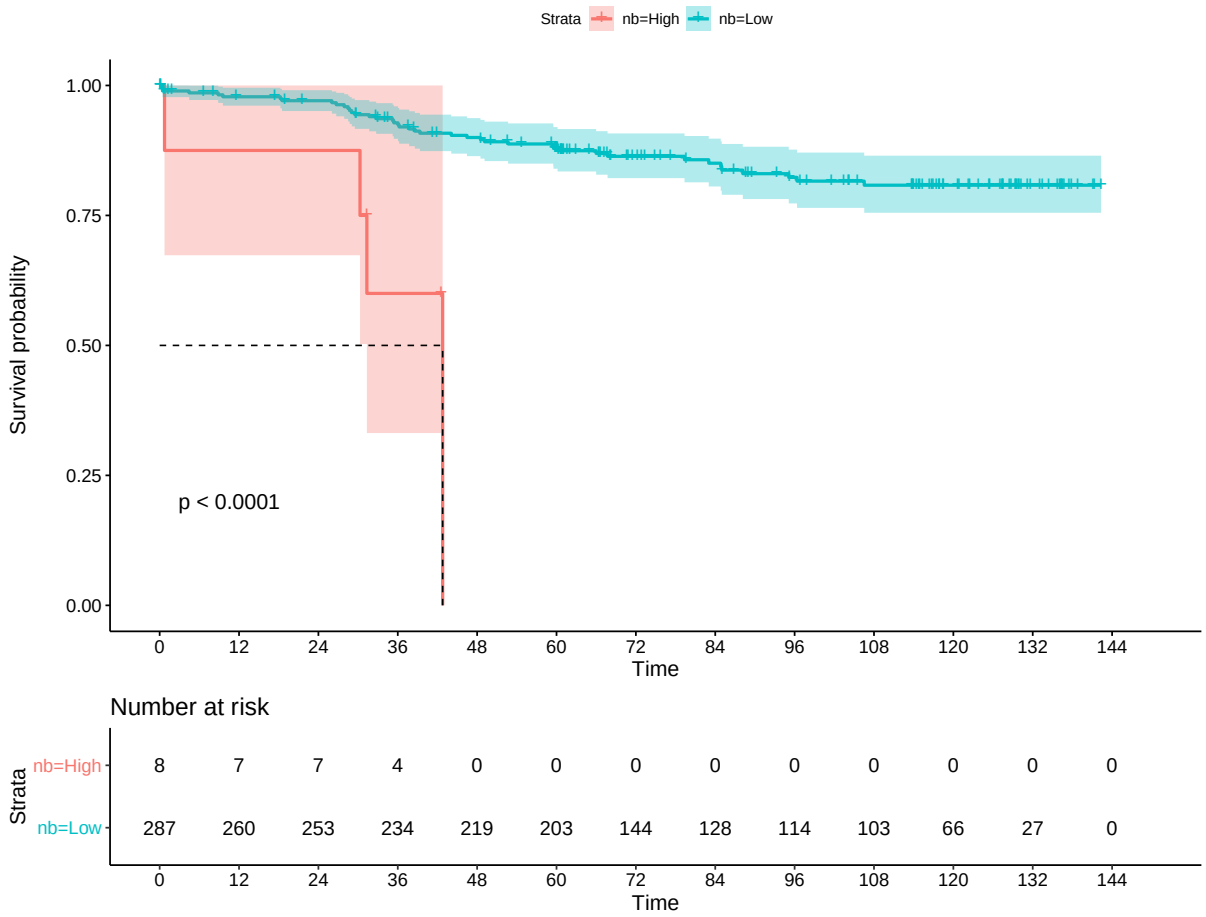


Figure 3.49: KM plot of survival probability (y-axis) versus time (x-axis) for NB with oversampling. NB here can identify patients with lower survival chance with a p-value <0.0001 in log-rank test while keeping the low risk group above 0.8 survival chance.

		Predicted	
		Low Risk	High Risk
Actual	Censored	245	4
	Dead	42	4

Table 3.11: Confusion matrix for NB with oversampling

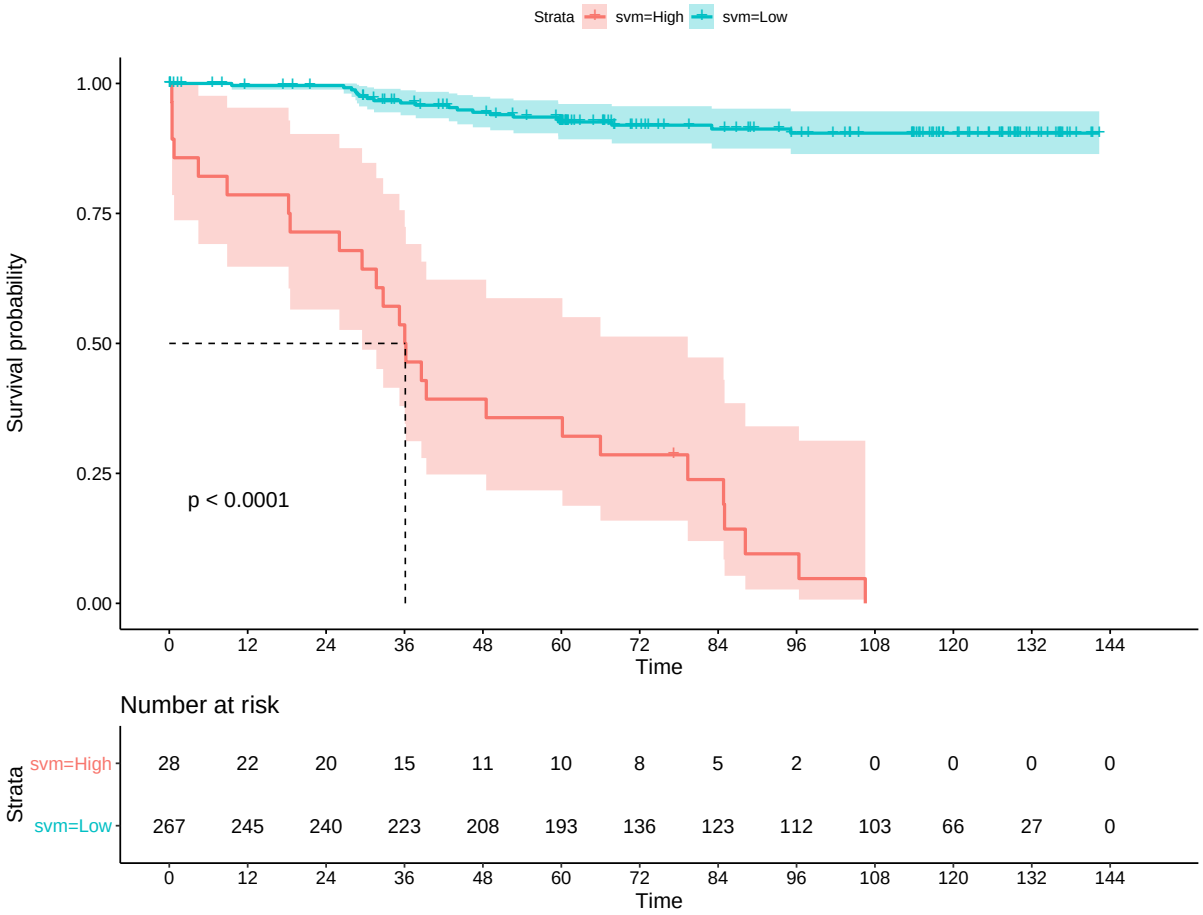


Figure 3.50: KM plot of survival probability (y-axis) versus time (x-axis) for SVM with oversampling. SVM can identify patients with lower survival chance with a median of time around 36 months and a p-value <0.0001 in log-rank test while keeping the low risk group above 0.9 survival chance.

		Predicted	
		Low Risk	High Risk
Actual	Censored	247	2
	Dead	20	26

Table 3.12: Confusion matrix for SVM with oversampling

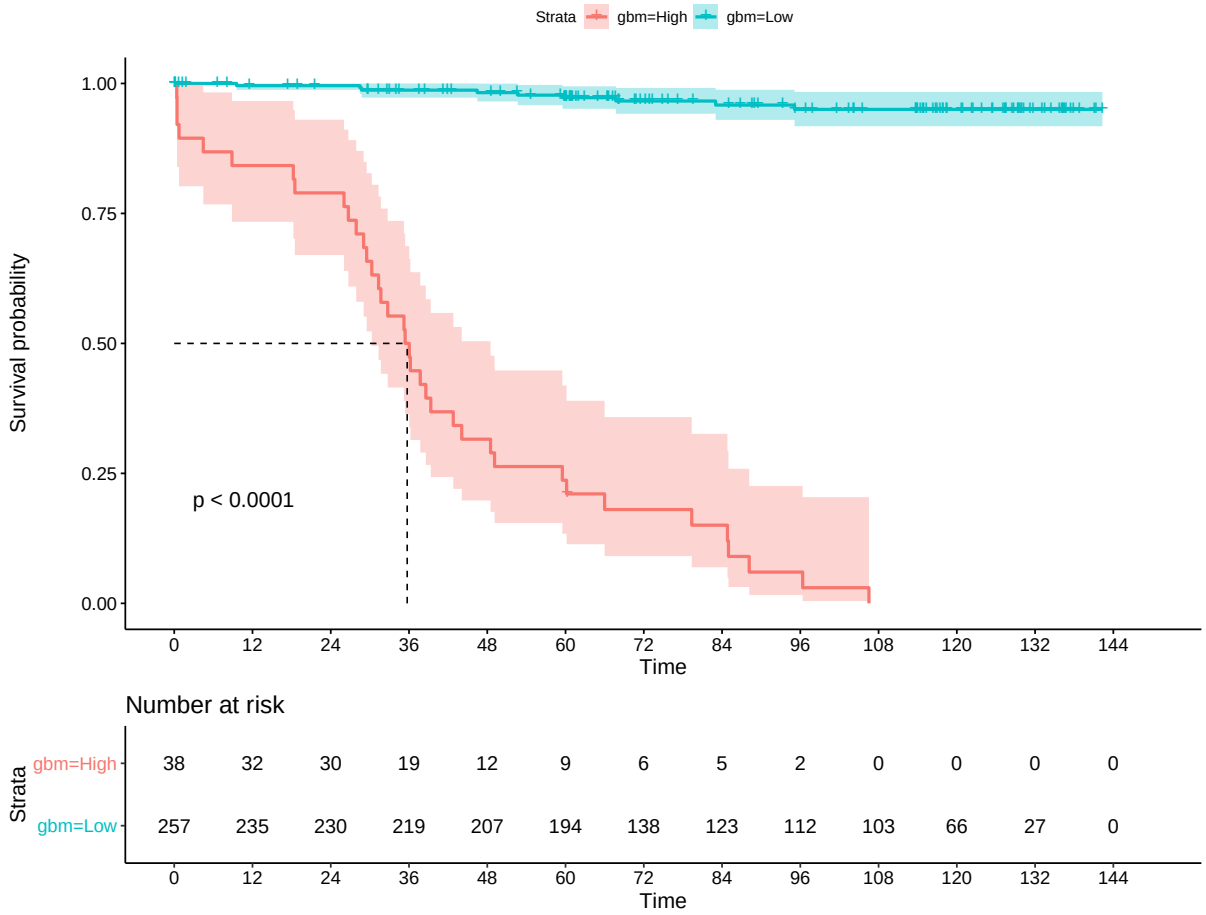


Figure 3.51: KM plot of survival probability (y-axis) versus time (x-axis) for gbm with oversampling. gbm can identify patients with lower survival chance with a median of time slightly less than 36 months and a p-value <0.0001 in log-rank test while keeping the low risk group above 0.95 survival chance.

		Predicted	
		Low Risk	High Risk
Actual	Censored	248	1
	Dead	9	37

Table 3.13: Confusion matrix for gbm with oversampling

3.4 Discussion

The density analysis is based on several assumptions. First, the slides are regarded as representative projections from three-dimensional cancer block section to two-dimensional slides, despite information loss. It is because a three-dimensional model is currently unavailable under

technological constraints; orthogonal cutting for a deeper exploration of the spatial structure is disallowed. Second, differences between cross-sectional and longitudinal cuts should not significantly affect the results. Third, slides from the same section should show similar density patterns and the variations across the slides should not substantially affect the results.

In section 3.2.1.9, it is interesting to see that the regularisation models trained with the original dataset do not work well with the data in general for the Edinburgh validation cohort, notwithstanding that the training dataset also comes from Edinburgh. The prediction performances of these models for the Edinburgh validation cohort, in terms of C-index, are no better or slightly better than random guessing. In contrast, these models work better for the Japanese validation cohort, especially for lasso and elastic net. In particular, the grouping effect helps elastic net to perform slightly better than lasso.

It is uncertain if oversampling can help improve the predictability for regularised models. Contrasting results are shown in Edinburgh and Japanese validation cohorts. Since one of the drawbacks of SMOTE oversampling is that it introduces additional artificial entries into the training dataset. This may cause regularised models to wrongly capture these noises as signals. As a result, some models perform worse after oversampling.

Given that the unsatisfactory results in regularisation, it is suspected that there are non-linearity. The non-linearity may be solved by adding interaction and polynomial terms into the regularised models. However, given the high dimensionality, it further complicates the problem if extra variables are added by trial and error.

Therefore, machine learning techniques are introduced to solve the non-linearity. It is found in the partial plots that RF and gbm can better capture the signal than those regularised models, while the latter outperforms the former. It is because the boosting property allows gbm to perform better in this imbalance dataset. Also, gbm can generalise for both validation cohorts, while regularised models cannot.

According to the variable importance of gbm, it highlights the proximity information of macrophage infiltrates around lymphocytic infiltrates. It supports further biological investigations since there is not much literature regarding their interaction.

For classification, none of the classification rules are able to perform consistently across the totally unseen Edinburgh, Japanese validation cohorts and the oversampled validation dataset. However, SVM and gbm both work well with the SMOTE oversampled training and validation cohorts. It suggests the potential of having a larger dataset.

For variable importance in NB, SVM and gbm, they mostly agree with the feature selection in the

survival analysis that the proximity information of macrophage infiltrates around lymphocytic infiltrates play a crucial role in classifying the high risk and low risk group. In particular, NB has only one predictor related to this. In addition, the latter two models highlight the importance of CD3 in CT. Surprisingly, the variables directly related to TBs do not appear in any of these models.

Oversampling can help improve this classification problem. Owing to the imbalance of the data, models simply predict all subjects as low risk group since most of the entries in the training dataset are censored. This is based on the assumption that censored patients are relatively lower risk than the dead patients. In fact, by oversampling the already oversampled training dataset in section 3.3.5.2 into a more balanced dataset, NB predictions on the oversampled validation dataset can be improved. Nevertheless, it still does not perform as well as SVM and gbm.

Given that SVM shares similar nature with discriminant analysis in reducing dimensionality, Penalised Linear Discriminant Analysis (PLDA), quadratic discriminant analysis (QDA) and Flexible Discriminant Analysis (FDA) were carried out for the classification problem, but to no avail. This also points to the non-linearity. For these reasons, discriminant analysis is not presented here. From the results, SVM with radial basis kernel suits better with the data structure here.

This study illustrates that SVM and gbm can tolerate interaction between predictors and collinearity. Nonetheless, interpreting these models are not as straightforward as the non-machine learning models.

CHAPTER FOUR

CONCLUSION

This chapter summarises the findings in this study and outlines the future works that can be done.

This study undergoes survival and classification analyses on the data of CRC patients, which contain clinicopathological and spatial density information. The latter was projected from a three-dimensional spatial distribution into two-dimensional. It is because immunofluorescence staining with automated image analysis can only project the spatial distribution from a three-dimensional tumour section to a two-dimensional slide. Thus, substantial spatial information is inevitably lost given current technological constraint.

The data in this study is of high dimensionality, collinearity and imbalanced responses. High dimensionality and collinearity are partly caused by trying to capture as many details of the proximity as possible. In the cases of proximity information, the density of macrophage infiltrates around lymphocytic infiltrates is subdivided into every 50 μm. This entails to collinearity since the density at $0-50\mu m$ is highly correlated to that of $50-100\mu m$ and so on. Meanwhile, this creates extra variables in order to capture the distribution into the attributes. The imbalance of the data makes the problem more complex since censoring is regarded as non-informative in Cox's models. What makes it worse is that these informative deaths may include outliers. This leads to data insufficiency for modelling.

To ease the high dimensionality and collinearity, regularised and machine learning models were constructed for survival analysis, where the latter ones outperform the former ones. This suggests non-linearity of the data, on top of dimensionality and collinearity. gbm survival model has the best predictability, in terms of C-index. In view of the non-linearity, machine learning techniques were also used for the classification analysis. Among these, SVM and gbm, which can tolerate interaction and collinearity between predictors, perform satisfactorily. In addition, NB model performs unexpectedly well in the unseen validation cohorts with only one predictor.

Hence, there is a promising potential of investigating how the interaction of macrophage and lymphocytic infiltrates influences CRC cancer.

The heterogeneity across the WTS may undermine the significant effects of selected features. For example, heavy clustering of certain cells at one location and no clustering in the other within CT cannot be reflected in corresponding predictor since this characteristic is averaged out across the region. This heterogeneity should be addressed but the complexity of this problem is beyond the scope of this study. Although the heterogeneity across WTS cannot fully explained under the given data framework, the final survival and classification models show only CD3 in CT and CD68 $100-200\mu m$ from CD3/CD8 are vital predictors. These suggest that density data from a closer distance are less important which merits further biological justification. Another direction of future work is a spatial analysis into the proximity of the macrophage infiltrates around lymphocytic infiltrates. Given that the information of their Cartesian coordinates is available, they can be analysed through spatial point pattern. In fact, Jones-Todd et al. proposed a parent-daughter point process to describe the spatial point pattern of the tumour and stroma cells within tissue samples [37]. They fitted a void process and two Neyman-Scott point process (NSPP), i.e. a Thomas and Matérn process, to explain the point pattern as a simplified model, first based on the assumption of homogeneous Poisson Process and adjusted for the edge effect thereafter. Nonetheless, they pointed out that the complexity of modelling through spatial point pattern analysis lies on the heterogeneity of the spatial structure in the tissue sections. This study can be extended to the macrophage infiltration.

Given the imbalanced data, it can be improved with more observations, as shown in the projection. However, it is often uneasy due to the expensive extraction of biological spatial data. If extra funding allows to recruit more patients, a further refinement of the classification rule developed here can be done to sub-categorise the high and low risk group;or else at least a power testing can take these findings a step further.

In view of the difference in Edinburgh and Japanese cohorts, swapping the cohorts for training and validation and carrying out a comparison may help bring the investigation further and to a greater depth. This may be due to dietary or genetic difference in Edinburgh and Japan. The difference of longitudinal and cross-sectional cut between the Edinburgh and Japanese cohorts may also merit further study. It is a blessing and a curse that there are three separate cohorts in this study. On one hand, the validation can be done on completely unseen data. On the other, it is worth considering to train the model with a larger proportion of the data.

To conclude, this study provides novel insights into the interaction of lymphocytic and macrophage infiltrates. Further works should be done on developing a more precise measure

of their interactions. It is hoped that this study can be generalised to other types of cancers and automated image analysis for survival can be further developed and help with less patient suffering in near future.

APPENDIX A

R SCRIPT AND CONSOLE OUTPUT

A.1 R Script and Data file

A list of R Script enclosed with this dissertation and data files used is as follows:

File name	Description
Data file	
All data All cohorts.csv	► Spatial density and clinicopathological data for all patients
Proximity All Cohorts.csv	► Proximity data for all patients
R Script file	
data input.R	► Load all libraries required and input the data.
rosesmoteover.R	► Create the oversampled data.
glmnet.R	► Create the regularised models on the original dataset.
glmnet oversample.R	► Create the regularised models on the oversampled dataset.
glmnetboot.R	► Generate the bootstrapping results for regularised models.
RFsrc.R	► Create the RF survival models.
gbmsrc.R	► Create the gbm survival models.
pT.R	► Generate the survival output for pT Stage.
NB.R	► Create the output of NB classification models.
SVM.R	► Create the output of SVM classification models.
gbm.R	► Create the output of gbm classification models.

Table A.1: A list of R Script and data files enclosed

We followed Nearchou *et al.* [3] in separating the training and validation data. That is the Edinburgh cohort in 2002 adopted as training dataset and the other two as validation datasets, as shown in figure 1.1. Due to the class imbalance, the minority, i.e. the dead patients, in the training dataset were over-sampled to with class probability of 0.5 via the `DMwR` and `ROSE` package in `R` [38, 39] (see section 2.4). Their model performances were compared and were checked against with the Cross Validation (CV) error if oversampling could help improve accuracy.

For survival analysis, regularised models (see section 2.2), RF and gbm (see section 2.3 were implemented to reduce the high dimensionality and undergo feature selection via the `glmnet`, `randomForestSRC` and `gbm` packages in `R` respectively [18, 23, 25, 29, 30]. The package `ggRandomForests` in `R` was used for visualisation of the RF survival models [40]. Regularised models were built using ten-fold CV with the `cv.glmnet` function in the `glmnet` package [18]. Dummy variables were created to fit the `glmnet` package in `R`, which cannot cope with factor variables [18, 23]. A bootstrapping analysis was conducted to ascertain the stability of these

regularised models. Variable importance of the RF and gbm were examined to compare the difference of results between machine learning and regularised models, as well as for any non-linearity existed. Predictability assessments were generated via the `Hmisc` package in `R` [41].

The analysis was then reduced into a classification problem. Given the certain non-linearity found in the previous stage, RF, gbm, NB and SVM (see section 2.3) were implemented to predict the binary DSD response via the `caret` package, which in turns calls the `randomForest`, `gbm`, `kernlab` and `naivebayes` packages in `R` [30, 36, 42–44]. five-fold CV with SMOTE was executed repeated for ten times since machine learning techniques usually require a larger fold sample size [34]. SMOTE prevents these techniques from tilting towards the censored observations, otherwise these algorithms would simply classify most of observations as censored patients. Based on these algorithms, classification systems to grade the CRC patients into high and low risks group were created. For the assessment of their performance, the `survival` and `survminer` packages in `R` are used for generating the KM plots and log-rank tests were produced against the validation cohorts [45, 46].

A missing entry for differentiation in the Japanese validation cohort is replaced by the mode, i.e. moderately differentiated for validation.

This dissertation is compiled by LaTeX.

A.2 Console Output

```
> cr
                                      nr Active.Coefficients.r
29                      CD68.CD163.in.CT           2.643910e-01
30                CD68.CD163.IN.IM.AND.CT           2.144527e-01
28                      CD68.CD163.in.IM           1.498931e-01
123                                Diff2           8.623363e-02
118                                 Age2           8.428900e-02
124                                Diff3          -8.279393e-02
34           CD68pCD163n.overCD163.in.IM          -6.472537e-02
8                          CD3overCD8inCT           5.722372e-02
121                                Site2          -5.637517e-02
16                            CD8.0.50.TB           5.616602e-02
> cr.oversample
                                      nr Active.Coefficients.r
29                      CD68.CD163.in.CT           2.643910e-01
30                CD68.CD163.IN.IM.AND.CT           2.144527e-01
28                      CD68.CD163.in.IM           1.498931e-01
123                                Diff2           8.623363e-02
118                                 Age2           8.428900e-02
124                                Diff3          -8.279393e-02
34           CD68pCD163n.overCD163.in.IM          -6.472537e-02
```

```
8                         CD3overCD8inCT          5.722372e-02
121                               Site2         -5.637517e-02
16                          CD8.0.50.TB          5.616602e-02
```

Listing A.1: Top 10 influential coefficients of ridge regression extract

```
> cl
                      nl Active.Coefficients.l
10                  Age2           0.601191068
3         CD3overCD8inCT           0.599545464
6            CD8.0.50.TB           0.509493965
2         CD3overCD8inIM          -0.301974125
8    CD68pCD163n.0.50.TB           0.184422396
7      CD163.PER.TB.0.50          -0.123006282
1                   CD8CT          -0.114207649
9      CD68..200.250.CD8          -0.099725138
11                 Site2          -0.038951742
4               TBNumber           0.009135314
5  CD3.0.50.TB.TB.number          -0.006717953
> cl.oversample
                                     nl Active.Coefficients.l
13        CD68pCD163n.overCD163.in.IM         -4.324645e+01
7                          CD8.0.50.TB         -1.268550e+01
32               CD68pCD163n.0.50.CD3         -8.831854e+00
1                                CD3IM          8.715028e+00
8             CD3CD8.0.50.TB.TB.Number         -5.993293e+00
12                    CD68pCD163n.IMCT          4.607770e+00
27                    CD68.100.150.CD8          4.297900e+00
15                    CD163.CD68.in.CT          3.508285e+00
24                    CD68..50.100.CD3          3.498057e+00
23                      CD68..0.50.CD3         -3.115038e+00
16                           CD163IMCT         -3.108023e+00
14                    CD163.CD68.in.IM         -3.014526e+00
4                CD3overCD8inIMandCT2         -2.355891e+00
5                          TB.Density          1.744236e+00
25                    CD68.100.150.CD3          1.742834e+00
22   Area.of.Whole.Slide.Image..mm.2.         -1.680369e+00
11                     CD68pCD163n..IM          1.589893e+00
19                       CD163.0.50.TB          1.523754e+00
18                    CD68.PER.TB.0.50         -1.512387e+00
31                   CD163.200.250.CD8          1.227718e+00
9                           CD68.in.IM          1.143956e+00
29                   CD163.200.250.CD3         -9.878216e-01
10                         CD163.in.IM         -4.250148e-01
17                       CD68..0.50.TB         -3.645799e-01
41                               EMLVI2         -2.521141e-01
21 CD163andCD68pCD163n.WITHIN.0.50.TB          2.327500e-01
2                                CD8IM          2.271455e-01
3                     CD3overCD8inCT         -1.862022e-01
35                               Age3          1.768483e-01
20             CD68.CD163.PER.TB.0.50         -1.514547e-01
```

```
37                              Site2         -1.483369e-01
6               CD3.0.50.TB.TB.number         -1.399847e-01
28                  CD163.100.150.CD3         -1.374213e-01
38                              Site3         -1.008610e-01
33                               Sex2          8.992729e-02
36                                pT2          6.131133e-02
40                              Diff3          4.708336e-02
39                              Diff2          4.562895e-02
34                               Age2         -3.423864e-02
26                  CD68.150.200.CD3          3.706640e-04
30                  CD163.50.100.CD8          7.203637e-09
```

Listing A.2: Coefficients of lasso selected variables

```
> ce2
                     ne2 Active.Coefficients.e2
12                  Age2              0.60760545
4         CD3overCD8inCT              0.59272321
7            CD8.0.50.TB              0.50477425
3         CD3overCD8inIM             -0.31196971
9    CD68pCD163n.0.50.TB              0.18126877
8      CD163.PER.TB.0.50             -0.12092067
10     CD68..200.250.CD8             -0.11809309
1                  CD8CT             -0.10128242
13                 Site2             -0.07217409
5               TBNumber              0.02951450
11                  Sex2              0.02138272
2                 CD8WTS             -0.01922241
6  CD3.0.50.TB.TB.number             -0.01633434
> ce2.oversample
                                  ne2 Active.Coefficients.e2
21        CD68pCD163n.overCD163.in.IM           -52.436740952
1                                CD3IM             7.234295882
10                         CD8.0.50.TB            -6.772638250
17                  CD68.CD163.in.CT              6.446484076
47               CD68pCD163n.0.50.CD3            -5.188977613
23                   CD163.CD68.in.CT              4.023702634
12                     CD3CD8.0.50.TB            -3.356281793
36                   CD68..50.100.CD3              3.278496514
25                      CD68..0.50.TB            -3.088476670
13           CD3CD8.0.50.TB.TB.Number            -3.008983481
18                    CD68pCD163n..IM              2.944693647
35                     CD68..0.50.CD3            -2.894628799
22                   CD163.CD68.in.IM            -2.645307572
11               CD8.0.50.TB.TB.number            -2.507987489
20                   CD68pCD163n.IMCT              2.469531561
37                   CD68.100.150.CD3              2.092826765
38                     CD68..0.50.CD8            -2.085310208
28                      CD163.0.50.TB              2.040831549
27                     CD68..0.100.TB              2.005167112
34   Area.of.Whole.Slide.Image..mm.2.            -1.985801648
```

```
40                    CD68.100.150.CD8            1.910112379
3                                CD8IM            1.902335718
24                           CD163IMCT           -1.800045907
7                           TB.Density            1.687181859
39                    CD68..50.100.CD8            1.580314414
5                 CD3overCD8inIMandCT2           -1.193196962
8                CD3.0.50.TB.TB.number           -1.154448960
15                         CD163.in.IM           -1.054182015
19                      CD68pCD163n..CT            1.007512665
46                   CD163.200.250.CD8            0.995774576
48               CD68pCD163n.0.100.CD3           -0.985791081
31          CD68.CD163.WITHIN.0.100.TB            0.962714482
43                   CD163.100.150.CD3           -0.919869507
41                    CD68.150.200.CD8            0.847142486
44                   CD163.200.250.CD3           -0.837223354
16                         CD163.in.CT           -0.787119144
4                       CD3overCD8inCT           -0.683672375
6                              TBNumber           -0.573907821
26                    CD68.PER.TB.0.50           -0.539229759
29                  CD163.PER.TB.0.100           -0.476902198
14                           CD68.in.IM            0.475668945
30              CD68.CD163.PER.TB.0.50           -0.447320587
2                                CD3WTS            0.341087283
33                                EMLVI           -0.297484148
9               CD3.0.100.TB.TB.number           -0.282699000
45                   CD163.50.100.CD8            0.263224825
58                               EMLVI2           -0.231278755
52                                 Age3            0.190937221
54                                Site2           -0.160518799
32 CD163andCD68pCD163n.WITHIN.0.50.TB            0.129798148
55                                Site3           -0.117224973
50                                 Sex2            0.114714643
42                   CD68..200.250.CD8            0.099214571
53                                  pT2            0.074677064
56                                Diff2            0.059537543
57                                Diff3            0.039021604
51                                 Age2           -0.031536627
49                CD68pCD163n.0.50.CD8           -0.006610829
```

Listing A.3: Coefficients of elastic net ($\alpha = 0.9$) selected variables

```
> cal
               nal Active.Coefficients.al
1   CD3overCD8inCT              0.4805789
2      CD8.0.50.TB              0.4555667
3 CD68.CD163.in.CT              2.4751353
4             Age2              0.5994154
> cal.oversample
                            nal Active.Coefficients.al
12 CD68pCD163n.overCD163.in.IMCT          -31.869711966
8                CD68.CD163.in.IM           31.037602501
```

```
11   CD68pCD163n.overCD163.in.IM          -10.027208712
9               CD68.CD163.in.CT            6.455665196
4                    CD8.0.50.TB           -3.780631261
1                          CD3IM            2.971673047
10              CD68pCD163n.IMCT            2.962596433
24          CD68pCD163n.0.50.CD3           -2.887513890
21                CD68..0.50.CD8           -2.459625732
19              CD68..50.100.CD3            2.159462984
20              CD68.100.150.CD3            1.210696771
6       CD3CD8.0.50.TB.TB.Number           -1.076429641
16                 CD163.0.50.TB            0.990705986
5          CD8.0.50.TB.TB.number           -0.919394984
14              CD163.CD68.in.CT            0.832925032
7                    CD163.in.CT           -0.682990101
15                CD68..0.50.TB           -0.569906822
18                         EMLVI           -0.540327350
13              CD163.CD68.in.IM           -0.486718112
3                    TB.Density            0.375236463
26                         Age3            0.354324872
22              CD68..50.100.CD8            0.335817640
31                        EMLVI2           -0.280435755
17            CD163.PER.TB.0.100           -0.266782391
25                         Sex2            0.206077965
28                        Site2           -0.181650007
29                        Site3           -0.133517525
2               CD3overCD8inCT            -0.120948084
27                          pT2            0.050245595
23            CD163.200.250.CD3           -0.027693894
30                         Diff2            0.004144669
```

Listing A.4: Coefficients of adaptive lasso ($\gamma = 1$) selected variables

```
> cal2.oversample
                            nal2 Active.Coefficients.al2
6               CD68.CD163.in.IM              32.21094041
10 CD68pCD163n.overCD163.in.IMCT             -27.15571444
9    CD68pCD163n.overCD163.in.IM             -11.58243139
7               CD68.CD163.in.CT               6.15138031
3                    CD8.0.50.TB              -2.56495378
18          CD68pCD163n.0.50.CD3              -2.36395828
8               CD68pCD163n.IMCT               1.93144358
1                          CD3IM               1.91537965
15              CD68..50.100.CD3               1.87625669
17                CD68..0.50.CD8              -1.77700342
4          CD8.0.50.TB.TB.number              -1.37812287
16              CD68.100.150.CD3               1.11036467
14            CD163.PER.TB.0.100              -0.57941537
13                 CD163.0.50.TB               0.56827304
12                CD68..0.50.TB              -0.40443643
5                   CD163.in.CT              -0.38585066
20                         Age3               0.32641023
```

```
2                      TB.Density              0.22354208
19                           Sex2              0.18357874
11               CD163.CD68.in.CT              0.13993132
22                           Site2            -0.10596627
23                           Site3            -0.09812822
24                          EMLVI2            -0.07968267
21                             pT2              0.07586564
> cal3.oversample
                           nal3 Active.Coefficients.al3
3 CD68pCD163n.overCD163.in.IM               -7.365046
2             CD68.CD163.in.CT                6.664329
1             CD68.CD163.in.IM                5.043118
```

Listing A.5: Coefficients of adaptive lasso ($\gamma = 2, 3$) selected variables

```
% lasso bootstrap probability of zero < 5%
% original
> l30
                   CD68.100.150.CD3
                              0.000
% oversample
> head(l20, 21)
                  CD68pCD163n..CT                  CD68..50.100.CD3
                            0.000                             0.000
                             Age2                              Age3
                            0.000                             0.000
                              pT2          CD3CD8.0.50.TB.TB.Number
                            0.000                             0.001
                            Site2                             Diff3
                            0.001                             0.001
            CD68pCD163n.0.50.CD3                              Sex2
                            0.003                             0.003
                      TB.Density       CD68pCD163n.overCD163.in.IM
                            0.008                             0.013
Area.of.Whole.Slide.Image..mm.2.                              CD3IM
                            0.021                             0.022
                   CD68..0.50.TB                     CD163.0.50.TB
                            0.025                             0.025
                CD68.CD163.in.CT                  CD163.CD68.in.IM
                            0.032                             0.032
                           EMLVI2                       CD163.in.CT
                            0.033                             0.042
           CD8.0.50.TB.TB.number
                            0.043
```

Listing A.6: Probability of being zero for variables in lasso bootstrap

```
% elastic net bootstrap probability of zero < 5%
```

```
% original
> e30
                   CD68.100.150.CD3
                              0.000
% oversample
> head(e20, 44)
                              CD3IM                              CD3CT
                              0.000                              0.000
                         CD8.0.50.TB            CD8.0.50.TB.TB.number
                              0.000                              0.000
                       CD3CD8.0.50.TB          CD3CD8.0.50.TB.TB.Number
                              0.000                              0.000
                           CD68.in.IM                     CD163.in.CT
                              0.000                              0.000
                      CD68pCD163n..CT                  CD68pCD163n.IMCT
                              0.000                              0.000
       CD68pCD163n.overCD163.in.IM                  CD163.CD68.in.IM
                              0.000                              0.000
                        CD68..0.50.TB                    CD163.0.50.TB
                              0.000                              0.000
  Area.of.Whole.Slide.Image..mm.2.                    CD68..0.50.CD3
                              0.000                              0.000
                    CD68..50.100.CD3                  CD68.100.150.CD3
                              0.000                              0.000
                      CD68..0.50.CD8                 CD163.200.250.CD3
                              0.000                              0.000
               CD68pCD163n.0.50.CD3           CD68pCD163n.100.150.CD3
                              0.000                              0.000
               CD68pCD163n.0.50.CD8             CD68pCD163n.0.100.CD8
                              0.000                              0.000
                               Sex2                               Age2
                              0.000                              0.000
                               Age3                                pT2
                              0.000                              0.000
                              Site2                              Diff3
                              0.000                              0.000
                              CD8CT             CD3.0.50.TB.TB.number
                              0.001                              0.001
             CD3.0.100.TB.TB.number                        TB.Density
                              0.003                              0.004
                   CD68.PER.TB.0.50                           EMLVI2
                              0.006                              0.007
                   CD68.CD163.in.CT           CD68pCD163n.50.100.CD3
                              0.008                              0.016
                  CD163.200.250.CD8                         CD163IMCT
                              0.020                              0.033
              CD68pCD163n.0.150.CD8                CD68pCD163n..IM
                              0.037                              0.040
CD163andCD68pCD163n.WITHIN.0.50.TB             CD68pCD163n.0.100.CD3
                              0.043                              0.048
```

Listing A.7: Probability of being zero for variables in elastic net bootstrap

A.2.1 Random Forest (RF) Survival Model

```
> print(v.obj5)
                         Sample size: 113
                  Number of deaths: 15
                   Number of trees: 5000
         Forest terminal node size: 5
      Average no. of terminal nodes: 6.3662
No. of variables tried at each split: 12
             Total no. of variables: 122
      Resampling used to grow trees: swor
   Resample size used to grow trees: 71
                           Analysis: RSF
                             Family: surv
                   Splitting rule: logrank *random*
      Number of random split points: 10
                         Error rate: 15.23%
> v.obj$importance
                              CD3IM                              CD3CT
                       1.298597e-03                       4.550463e-03
                             CD3WTS                              CD8IM
                       3.527221e-03                       1.333233e-04
                              CD8CT                             CD8WTS
                       2.971816e-03                       7.287496e-04
                         CD3overCD8inIM                     CD3overCD8inCT
                       6.296773e-05                       1.401478e-04
                  CD3overCD8inIMandCT2                         TBNumber
                       6.299454e-05                      -2.312162e-04
                        TB.Density                        CD3.0.50.TB
                       2.165115e-04                       5.923494e-04
                CD3.0.50.TB.TB.number                        CD3.0.100
                       1.052420e-03                       9.919178e-04
               CD3.0.100.TB.TB.number                        CD8.0.50.TB
                       1.296735e-03                       3.874520e-04
                CD8.0.50.TB.TB.number                        CD8.0.100
                       6.611162e-04                       1.987601e-04
                  CD8.0.100.TB.number                     CD3CD8.0.50.TB
                       9.653845e-04                       4.339899e-04
             CD3CD8.0.50.TB.TB.Number                     CD3CD8.0.100.TB
                       1.603986e-03                       8.181651e-04
```

Listing A.8: RF survival model console output

A.2.2 Gradient Boosting Survival Model

```
> summary(bfitt)
var                                                     rel.inf
CD68.100.150.CD3                                CD68.100.150.CD3
    19.707309980
```

```
CD68.150.200.CD8                                    CD68.150.200.CD8
    9.898729561
CD3CT                                                         CD3CT
    9.503525746
CD68pCD163n.100.150.CD3                      CD68pCD163n.100.150.CD3
    4.346269980
CD163.200.250.CD3                                CD163.200.250.CD3
    4.110588092
CD68pCD163n.overCD163.in.CT                CD68pCD163n.overCD163.in.CT
    4.031621214
CD68.150.200.CD3                                  CD68.150.200.CD3
    3.952658332
CD3CD8.0.100.TB..TB.Number                  CD3CD8.0.100.TB..TB.Number
    3.222509088
CD68pCD163n..CT                                    CD68pCD163n..CT
    3.165678480
CD68.CD163.IN.IM.AND.CT                    CD68.CD163.IN.IM.AND.CT
    3.131575724
Area.of.Whole.Slide.Image..mm.2.       Area.of.Whole.Slide.Image..mm.2.
    2.640035308
CD68pCD163n.0.50.CD8                         CD68pCD163n.0.50.CD8
    2.324027600
CD163.100.150.CD8                               CD163.100.150.CD8
    1.674590360
CD68pCD163n.0.100.CD3                       CD68pCD163n.0.100.CD3
    1.631056185
Age                                                           Age
    1.512958672
CD163.PER.TB.0.100                             CD163.PER.TB.0.100
    1.442232550
CD8.0.50.TB.TB.number                       CD8.0.50.TB.TB.number
    1.426331629
CD163.150.200.CD3                               CD163.150.200.CD3
    1.346964568
CD3overCD8inCT                                     CD3overCD8inCT
    1.335051890
CD68pCD163n..IM                                   CD68pCD163n..IM
    1.140073025
CD68.100.150.CD8                                 CD68.100.150.CD8
    1.128772018
TB.Density                                             TB.Density
    1.090432778
CD68..50.100.CD3                                 CD68..50.100.CD3
    1.080476633
CD68pCD163n.0.50.CD3                         CD68pCD163n.0.50.CD3
    1.051622144
```

Listing A.9: Gradient boosting survival model console output

A.2.3 Naïve Bayes (NB)

```
> nbFit
Naive Bayes

999 samples
  1 predictor
  2 classes: 'Censored', 'Died'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 10 times)
Summary of sample sizes: 799, 799, 800, 799, 799, 800, ...
Resampling results across tuning parameters:

  laplace  adjust  ROC        Sens      Spec
  0.0      0.0     0.8996933  0.814303  0.8446465
  0.5      0.5     0.8996933  0.814303  0.8446465
  1.0      1.0     0.8996933  0.814303  0.8446465

Tuning parameter 'usekernel' was held constant at a value of TRUE
ROC was used to select the optimal model using the largest value.
The final values used for the model were laplace = 0, usekernel = TRUE and
    adjust = 0.

> confusionMatrix(predict(nbFit), data4$DiseaseSpecificDeath)
Confusion Matrix and Statistics

          Reference
Prediction Censored Died
  Censored      411   77
  Died           93  418

               Accuracy : 0.8298
                 95% CI : (0.8051, 0.8526)
    No Information Rate : 0.5045
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6597
 Mcnemar's Test P-Value : 0.25

            Sensitivity : 0.8155
            Specificity : 0.8444
         Pos Pred Value : 0.8422
         Neg Pred Value : 0.8180
             Prevalence : 0.5045
         Detection Rate : 0.4114
   Detection Prevalence : 0.4885
      Balanced Accuracy : 0.8300

       'Positive' Class : Censored
```

Listing A.10: NB classification model console output

A.2.4 Support Vector Machines (SVM)

```
> svmFit <- train(DiseaseSpecificDeath ~ CD68.150.200.CD3+
                    CD68..200.250.CD3+
                    CD68pCD163n.150.200.CD3+
                    CD68.100.150.CD3+
                    CD3CT+
                    CD68pCD163n.200.250.CD3+
                    CD68pCD163n.100.150.CD3+
                    CD68.150.200.CD8+
                    CD68pCD163n.150.200.CD8,
                  data = data4,
                  method = "svmRadial",
                  trControl = cv_5_grid,
                  tuneLength = 8,
                  metric = "ROC")
> svmFit
Support Vector Machines with Radial Basis Function Kernel

999 samples
  9 predictor
  2 classes: 'Censored', 'Died'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 10 times)
Summary of sample sizes: 799, 799, 800, 799, 799, 799, ...
Resampling results across tuning parameters:

  C      ROC        Sens       Spec
   0.25  0.9654847  0.8813604  0.8830303
   0.50  0.9712051  0.8986218  0.9101010
   1.00  0.9781551  0.9216139  0.9349495
   2.00  0.9822736  0.9301446  0.9442424
   4.00  0.9847753  0.9555584  0.9537374
   8.00  0.9873294  0.9658752  0.9628283
  16.00  0.9883635  0.9720218  0.9769697
  32.00  0.9906549  0.9799584  0.9842424

Tuning parameter 'sigma' was held constant at a value of 0.6662246
ROC was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.6662246 and C = 32.

> confusionMatrix(predict(svmFit), data4$DiseaseSpecificDeath)
Confusion Matrix and Statistics

          Reference
Prediction Censored Died
  Censored      498    2
  Died            6  493

               Accuracy : 0.992
                 95% CI : (0.9843, 0.9965)
    No Information Rate : 0.5045
```

```
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.984

 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.9881
            Specificity : 0.9960
         Pos Pred Value : 0.9960
         Neg Pred Value : 0.9880
             Prevalence : 0.5045
         Detection Rate : 0.4985
   Detection Prevalence : 0.5005
      Balanced Accuracy : 0.9920

       'Positive' Class : Censored
```

Listing A.11: SVM classification model console output

```
> confusionMatrix(predict(gbmFit.all), data4$DiseaseSpecificDeath)
Confusion Matrix and Statistics

          Reference
Prediction Censored Died
  Censored      504    0
  Died            0  495

               Accuracy : 1
                 95% CI : (0.9963, 1)
    No Information Rate : 0.5045
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5045
         Detection Rate : 0.5045
   Detection Prevalence : 0.5045
      Balanced Accuracy : 1.0000

       'Positive' Class : Censored

 > varImp(gbmFit.all)
gbm variable importance

  only 20 most important variables shown (out of 116)
```

```
                          Overall
CD3CT                     100.000
CD68.100.150.CD3           62.196
CD68.150.200.CD3           17.304
CD68.150.200.CD8           13.728
CD68..200.250.CD3          12.610
CD68pCD163n.150.200.CD3    10.002
CD3CD8.0.100.TB             9.473
CD3CD8.0.50.TB.TB.Number    7.987
CD163.PER.TB.0.100          7.242
CD68pCD163n.0.50.CD8        6.593
CD8.0.100.TB.number         5.805
CD8CT                       5.780
CD68pCD163n.200.250.CD3     5.621
CD3overCD8inCT              4.971
CD163..0.50.CD8             3.832
TB.Density                  3.610
CD3CD8.0.100.TB..TB.Number  3.221
CD163..0.200.CD8            3.070
CD68pCD163n.100.150.CD3     3.009
CD68.CD163.in.CT            2.643
```

Listing A.12: gbm classification model console output

```
# pT Stage
> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ pT,
+          data=v1s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    pT, data = v1s)

      N Observed Expected (O-E)^2/E (O-E)^2/V
pT=1 46        6     9.42      1.24      8.65
pT=2 10        5     1.58      7.38      8.65

 Chisq= 8.7  on 1 degrees of freedom, p= 0.003

> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ pT,
+          data=v2s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    pT, data = v2s)

      N Observed Expected (O-E)^2/E (O-E)^2/V
pT=1 54       11    11.95    0.0748     0.943
pT=2  7        2     1.05    0.8475     0.943

 Chisq= 0.9  on 1 degrees of freedom, p= 0.3
```

```
> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ pT,
+          data=data5Test)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    pT, data = data5Test)

       N Observed Expected (O-E)^2/E (O-E)^2/V
pT=1 236       32    38.08     0.971      5.66
pT=2  59       14     7.92     4.672      5.66

 Chisq= 5.7  on 1 degrees of freedom, p= 0.02

# NB
> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ nb,
+          data=v1s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    nb, data = v1s)

         N Observed Expected (O-E)^2/E (O-E)^2/V
nb=High 17        6     3.08      2.76      3.84
nb=Low  39        5     7.92      1.07      3.84

 Chisq= 3.8  on 1 degrees of freedom, p= 0.05

> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ nb,
+          data=v2s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    nb, data = v2s)

         N Observed Expected (O-E)^2/E (O-E)^2/V
nb=High  4        2    0.352     7.713      8.06
nb=Low  57       11   12.648     0.215      8.06

 Chisq= 8.1  on 1 degrees of freedom, p= 0.005

> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ nb,
+          data=data5Test)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    nb, data = data5Test)

          N Observed Expected (O-E)^2/E (O-E)^2/V
nb=High  18       10     1.88      35.2        37
nb=Low  277       36    44.12       1.5        37

 Chisq= 37  on 1 degrees of freedom, p= 1e-09

# SVM
```

```
> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ svm,
+          data=v1s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    svm, data = v1s)

          N Observed Expected (O-E)^2/E (O-E)^2/V
svm=High  6        3    0.918      4.73      5.17
svm=Low  50        8   10.082      0.43      5.17

 Chisq= 5.2  on 1 degrees of freedom, p= 0.02

> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ svm,
+          data=v2s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    svm, data = v2s)

          N Observed Expected (O-E)^2/E (O-E)^2/V
svm=High  2        1    0.268     1.994      2.05
svm=Low  59       12   12.732     0.042      2.05

 Chisq= 2.1  on 1 degrees of freedom, p= 0.2

> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ svm,
+          data=data5Test)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    svm, data = data5Test)

           N Observed Expected (O-E)^2/E (O-E)^2/V
svm=High  28       26     2.78     194.2       208
svm=Low  267       20    43.22      12.5       208

 Chisq= 208  on 1 degrees of freedom, p= <2e-16

#gbm
> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ gbm,
+          data=v1s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    gbm, data = v1s)

          N Observed Expected (O-E)^2/E (O-E)^2/V
gbm=High 17        6     3.08      2.76      3.84
gbm=Low  39        5     7.92      1.07      3.84

 Chisq= 3.8  on 1 degrees of freedom, p= 0.05

> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ gbm,
```

```
+          data=v2s)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    gbm, data = v2s)

          N Observed Expected (O-E)^2/E (O-E)^2/V
gbm=High 15        7     2.96      5.53      7.23
gbm=Low  46        6    10.04      1.63      7.23

 Chisq= 7.2  on 1 degrees of freedom, p= 0.007

> survdiff(Surv(DiseaseSpecificSurvival,as.numeric(DiseaseSpecificDeath)) ~ gbm,
+          data=data5Test)
Call:
survdiff(formula = Surv(DiseaseSpecificSurvival, as.numeric(DiseaseSpecificDeath
    )) ~
    gbm, data = data5Test)

           N Observed Expected (O-E)^2/E (O-E)^2/V
gbm=High  38       37     3.55     315.4       348
gbm=Low  257        9    42.45      26.4       348

 Chisq= 348  on 1 degrees of freedom, p= <2e-16
```

Listing A.13: Log-rank test console output for Edinburgh cohort, Japanese cohort and oversampled validation dataset

```
> sessionInfo()
R version 3.5.2 (2018-12-20)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)

Matrix products: default

locale:
[1] LC_COLLATE=English_United Kingdom.1252  LC_CTYPE=English_United Kingdom.1252
[3] LC_MONETARY=English_United Kingdom.1252 LC_NUMERIC=C
[5] LC_TIME=English_United Kingdom.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] klaR_0.6-14     MASS_7.3-51.1   e1071_1.7-0.1   kernlab_0.9-27  caret_
     6.0-81
 [6] lattice_0.20-38 survminer_0.4.3 ggpubr_0.2      magrittr_1.5    dplyr_
     0.8.0.1
[11] ggplot2_3.1.0   survival_2.43-3 gbm_2.1.5

loaded via a namespace (and not attached):
```

```
 [1] viridis_0.5.1        tidyr_0.8.2          viridisLite_0.3.0
 [4] splines_3.5.2        foreach_1.4.4        prodlim_2018.04.18
 [7] Formula_1.2-3        shiny_1.2.0          assertthat_0.2.0
[10] highr_0.7            stats4_3.5.2         latticeExtra_0.6-28
[13] yaml_2.2.0           ipred_0.9-8          pillar_1.3.1
[16] backports_1.1.3      glue_1.3.0           pROC_1.13.0
[19] digest_0.6.18        RColorBrewer_1.1-2   promises_1.0.1
[22] checkmate_1.9.0      colorspace_1.4-0     recipes_0.1.4
[25] cmprsk_2.2-7         htmltools_0.3.6      httpuv_1.4.5.1
[28] Matrix_1.2-15        plyr_1.8.4           timeDate_3043.102
[31] pkgconfig_2.0.2      broom_0.5.1          questionr_0.7.0
[34] purrr_0.3.0          xtable_1.8-3         scales_1.0.0
[37] km.ci_0.5-2          later_0.7.5          gower_0.1.2
[40] lava_1.6.5           KMsurv_0.1-5         randomForestSRC_2.9.0
[43] tibble_2.0.1         htmlTable_1.13.1     combinat_0.0-8
[46] generics_0.0.2       ellipsis_0.1.0       withr_2.1.2
[49] nnet_7.3-12          lazyeval_0.2.1       cli_1.0.1
[52] crayon_1.3.4         mime_0.6             fansi_0.4.0
[55] nlme_3.1-137         forcats_0.4.0        foreign_0.8-71
[58] class_7.3-15         tools_3.5.2          data.table_1.12.0
[61] stringr_1.4.0        munsell_0.5.0        cluster_2.0.7-1
[64] compiler_3.5.2       rlang_0.3.1          grid_3.5.2
[67] iterators_1.0.10     rstudioapi_0.9.0     htmlwidgets_1.3
[70] miniUI_0.1.1.1       labeling_0.3         base64enc_0.1-3
[73] gtable_0.2.0         ModelMetrics_1.2.2   codetools_0.2-16
[76] reshape2_1.4.3       R6_2.4.0             zoo_1.8-4
[79] gridExtra_2.3        lubridate_1.7.4      knitr_1.21
[82] survMisc_0.5.5       utf8_1.1.4           Hmisc_4.2-0
[85] stringi_1.3.1        parallel_3.5.2       Rcpp_1.0.0
[88] rpart_4.1-13         acepack_1.4.1        tidyselect_0.2.5
[91] xfun_0.5
```

Listing A.14: Session Information

APPENDIX B

SUPPLEMENTARY OUTPUTS

B.1 Proofs

Bias-variance trade-off decomposition:

Given $y = f(x^{'}) + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$ *and* $Var(\varepsilon) = \mathbb{E}(\varepsilon^2) = \sigma_\varepsilon^2$,

$$\begin{aligned}\mathbb{E}\left\{(\mathbf{y} - \hat{\mathbf{y}})^2\right\} &= \mathbb{E}\left\{\left(f(x^{'}) + \varepsilon - \hat{f}(x^{'}) + \mathbb{E}\hat{f}(x^{'}) - \mathbb{E}\hat{f}(x^{'})\right)^2\right\} \\ &= \mathbb{E}\left[f(x^{'}) - \mathbb{E}\hat{f}(x^{'})\right]^2 + \mathbb{E}(\varepsilon^2) + \mathbb{E}\left[\mathbb{E}\left(\hat{f}(x^{'}) - \hat{f}(x^{'})\right)^2\right] \\ &= \sigma_\varepsilon^2 + (Bias\{\hat{\mathbf{y}}\})^2 + Var\{\hat{\mathbf{y}}\}, \end{aligned} \tag{B.1}$$

OLS estimate:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ols} &= (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) \\ \frac{\partial \hat{\boldsymbol{\beta}}_{ols}}{\partial \beta} &= -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) \\ \frac{\partial^2 \hat{\boldsymbol{\beta}}_{ols}}{\partial \beta \partial \beta^T} &= 2\boldsymbol{X}^T\boldsymbol{X}\end{aligned}$$

As $\boldsymbol{X}^T\boldsymbol{X}$ is positive definite and first derivative is zero,

$$\begin{aligned}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) &= 0 \\ \hat{\beta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\end{aligned} \tag{B.2}$$

Variance of OLS estimator:

$$\begin{aligned}Var(\hat{\beta}_j) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\sigma_\varepsilon^2 I)(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma_\varepsilon^2.\end{aligned} \tag{B.3}$$

Ridge estimators in terms of OLS [47]:

Given an orthogonal design matrix $\mathbf{X}^T\mathbf{X} = \mathbf{I_{pp}} = (\mathbf{X}^T\mathbf{X})^{-1}$,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I_{pp}})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{I_{pp}} + \lambda\mathbf{I_{pp}})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (1+\lambda)^{-1}\mathbf{I_{pp}}\mathbf{X}^T\mathbf{y} \\ &= (1+\lambda)^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (1+\lambda)^{-1}\hat{\boldsymbol{\beta}}_{ols} \end{aligned} \tag{B.4}$$

The edf in ridge :

$$\begin{aligned}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I_{pp}})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I}_{pp})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y} \\ &= \sum_{j=1}^{p}\boldsymbol{u}_j\frac{d_j^2}{d_j^2+\lambda}\boldsymbol{u}_j^T\boldsymbol{y} \end{aligned} \tag{B.5}$$

where the u_j are the columns of $\boldsymbol{U}$ or the normalised principal component of $\boldsymbol{X}$, $\boldsymbol{U}$ and $\boldsymbol{D}$ are $n \times p$ orthogonal matrix , with the columns of $\boldsymbol{U}$ that spans the column space of $\boldsymbol{X}$, and $p \times p$ diagonal matrix with the diagonal values of $d_1 \geq ... \geq d_p \geq 0$ respectively [5]. $\boldsymbol{D}$ is also known as the singular values of $\boldsymbol{X}$.

Lasso estimators in terms of OLS [47]:

Given an orthogonal design matrix $\mathbf{X}^T\mathbf{X} = \mathbf{I_{pp}} = (\mathbf{X}^T\mathbf{X})^{-1}$,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{lasso} &= \underset{\beta}{\operatorname{argmin}}\left\{ \|\boldsymbol{y} - \sum_{\mathbf{j=1}}^{\mathbf{p}}\boldsymbol{x_j}\beta_{\mathbf{j}}\|^2 + \lambda\sum_{j=1}^{p}|\beta_j| \right\} \\ &= \underset{\beta}{\operatorname{argmin}}\left\{ \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\sum_{j=1}^{p}|\beta_j| \right\} \\ &\propto \underset{\beta}{\operatorname{argmin}}\left\{ -\hat{\boldsymbol{\beta}}_{ols}^T\boldsymbol{\beta} - \boldsymbol{\beta}^T\hat{\boldsymbol{\beta}}_{ols} + \boldsymbol{\beta}^T\boldsymbol{\beta} + \lambda\sum_{j=1}^{p}|\beta_j| \right\} \\ &= \sum_{j=1}^{p}\underset{\beta_j}{\operatorname{argmin}}\left\{ -2\hat{\beta}_j^{ols}\beta_j + \beta_j^2 + \lambda|\beta_j| \right\} \end{aligned}$$

Consider individual β_j,

$$\underset{\beta_j}{\operatorname{argmin}} \left\{ -2\hat{\beta}_j^{ols}\beta_j + \beta_j^2 + \lambda|\beta_j| \right\} = \begin{cases} \operatorname{argmin}_{\beta_j} \left\{ -2\hat{\beta}_j^{ols}\beta_j + \beta_j^2 + \lambda\beta_j \right\}, & \text{if } \beta_j \geq 0 \\ \operatorname{argmin}_{\beta_j} \left\{ -2\hat{\beta}_j^{ols}\beta_j + \beta_j^2 - \lambda\beta_j \right\}, & \text{if } \beta_j < 0 \end{cases} \tag{B.6}$$

By solving equation B.6,

$$\hat{\beta}_j^{lasso} = \begin{cases} \hat{\beta}_j^{ols} - \frac{\lambda}{2} & \text{if } \beta_j \geq 0 \\ \hat{\beta}_j^{ols} + \frac{\lambda}{2}, & \text{if } \beta_j < 0 \end{cases} \tag{B.7}$$

Combine the results in equation B.7,

$$\hat{\boldsymbol{\beta}}_{lasso} = sign(\hat{\boldsymbol{\beta}}_{ols})(|\hat{\boldsymbol{\beta}}_{ols}| - \lambda)_+ \tag{B.8}$$

SVM inner product [5]:

Define

$$\mathbf{M} = \frac{1}{\|\beta\|} \tag{B.9}$$

Then

$$\underset{\beta_0,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n,M}{\operatorname{argmax}} \mathbf{M} \text{ subject to } \begin{cases} \sum_{j=1}^p \beta_j^2 = 1, \\ \boldsymbol{y}_i \sum_{j=1}^p \mathbf{x_j}\beta_j \geq (1-\varepsilon_i), \\ \sum_{i=1}^n \varepsilon_i \leq \mathbf{C}, \end{cases} \tag{B.10}$$

Transform B.10 into,

$$\underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|\beta\|^2 + \mathbf{C}\sum_{i=1}^n \varepsilon_i \text{ subject to } \begin{cases} \sum_{j=1}^p \beta_j^2 = 1, \\ \boldsymbol{y}_i \sum_{j=1}^p \mathbf{x_j}\beta_j \geq (1-\varepsilon_i), \\ \sum_{i=1}^n \varepsilon_i \leq \mathbf{C}, \end{cases} \tag{B.11}$$

Restructure B.11 into Lagrange function,

$$L_p = \frac{1}{2}\|\beta\|^2 + \mathbf{C}\sum_{i=1}^{n}\varepsilon_i - \sum_{i}^{n}\alpha_i[y_i(x_i^T\beta) - (1-\varepsilon_i)] - \sum_{i=1}^{n}\mu_i\varepsilon_i, \tag{B.12}$$

where $\mu_i, \alpha_i \geq 0$ are variable constraints.

Minimise B.12 with respect to β and ε_i and assign corresponding derivatives to zero

$$\begin{aligned} \beta &= \sum_{i=1}^{n}\alpha_i y_i x_i, \\ \sum_{i=1}^{n}\alpha_i y_i &= 0, \\ \alpha_i &= \mathbf{C} - \mu_i. \end{aligned}$$

Solve the above,

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i}^{n}\sum_{i'}^{n}\alpha_i\alpha_{i'}y_i y_{i'}x_i^T x_{i'} \tag{B.13}$$

Re-express B.13

$$L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i}^{n}\sum_{i'}^{n}\alpha_i\alpha_{i'}y_i y_{i'}\langle x, x_i\rangle \tag{B.14}$$

Hence,

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle, \tag{B.15}$$

B.2 Tables

Variable	Description
Case No	► Patient case number
Area of Whole Slide Image (mm^2)	► Area of whole immunofluorescence staining slide image in mm^2
CD68 0-50 CD3, norm CD68 0-50 CD3, CD68 50-100 CD3 , CD68 0-100 CD3, norm CD68 0-100 CD3, CD68 100-150 CD3 , CD68 0-150 CD3, norm CD68 0-150 CD3, CD68 150-200 CD3, CD68 0-200 CD3, norm CD68 0-200 CD3, CD68 200-250 CD3, CD68 0-250 CD3, norm CD68 0-250 CD3, CD68 0-50 CD8, norm CD68 0-50 CD8, CD68 50-100 CD8 , CD68 0-100 CD8, norm CD68 0-100 CD8, CD68 100-150 CD8 , CD68 0-150 CD8, norm CD68 0-150 CD8, CD68 150-200 CD8, CD68 0-200 CD8, norm CD68 0-200 CD8, CD68 200-250 CD8, CD68 0-250 CD8, norm CD68 0-250 CD8, CD163 0-50 CD3 , norm CD163 0-50 CD3 , CD163 50-100 CD3 , CD163 0-100 CD3 , norm CD163 0-100 CD3 , CD163 100-150 CD3, CD163 0-150 CD3 , norm CD163 0-150 CD3 , CD163 150-200 CD3, CD163 0-200 CD3 , norm CD163 0-200 CD3 , CD163 200-250 CD3 , CD163 0-250 CD3 , norm CD163 0-250 CD3 , CD163 0-50 CD8, norm CD163 0-50 CD8, CD163 50-100 CD8 , CD163 0-100 CD8, norm CD163 0-100 CD8, CD163 100-150 CD8, CD163 0-150 CD8, norm CD163 0-150 CD8, CD163 150-200 CD8, CD163 0-200 CD8, norm CD163 0-200 CD8, CD163 200-250 CD8 , CD163 0-250 CD8, norm CD163 0-250 CD8, CD68pCD163n 0-50 CD3, norm CD68pCD163n 0-50 CD3, CD68pCD163n 50-100 CD3, CD68pCD163n 0-100 CD3, norm CD68pCD163n 0-100 CD3, CD68pCD163n 100-150 CD3, CD68pCD163n 0-150 CD3, norm CD68pCD163n 0-150 CD3, CD68pCD163n 150-200 CD3, CD68pCD163n 0-200 CD3, norm CD68pCD163n 0-200 CD3, CD68pCD163n 200-250 CD3, CD68pCD163n 0-250 CD3, norm CD68pCD163n 0-250 CD3, CD68pCD163n 0-50 CD8, norm CD68pCD163n 0-50 CD8, CD68pCD163n 50-100 CD8, CD68pCD163n 0-100 CD8, norm CD68pCD163n 0-100 CD8, CD68pCD163n 100-150 CD8, CD68pCD163n 0-150 CD8, norm CD68pCD163n 0-150 CD8, CD68pCD163n 150-200 CD8, CD68pCD163n 0-200 CD8, norm CD68pCD163n 0-200 CD8, CD68pCD163n 200-250 CD8, CD68pCD163n 0-250 CD8, norm CD68pCD163n 0-250 CD8	► Number or normalised (norm) number of macrophage (CD163 or CD68) infiltrates in vincity of lymphocytic (CD3 or CD8) from a specified distance (in μm) CD68pCD163n: CD68 positive and CD163 negative
DiseaseSpecificSurvival	► DSS time in months
DiseaseSpecificDeath	► Binary outcome for Disease Specific Death: 1 if died and 0 if censored

Table B.1: Variables description table of proximity variables

Variable	Description
CaseNo	► Patient case number
CD3IM, CD3CT, CD3WTS, CD8IM, CD8CT, CD8WTS, CD3overCD8inIM, CD3overCD8inCT, CD3overCD8inIMandCT2, CD3 0-50 TB, CD3 0-50 TB/TB number, CD3 0-100, CD3 0-100 TB/TB number, CD8 0-50 TB, CD8 0-50 TB/TB number, CD8 0-100, CD8 0-100/TB number, CD3CD8 0-50 TB, CD3CD8 0-50 TB/TB Number, CD3CD8 0-100 TB, CD3CD8 0-100 TB/ TB Number, CD68 in IM, CD68 in CT, CD163 in IM, CD163 in CT, CD68/CD163 in IM, CD68/CD163 in CT, CD68/CD163 IN IM AND CT , CD68pCD163n IM, CD68pCD163n CT, CD68pCD163n IMCT, CD68pCD163n overCD163 in IM, CD68pCD163n overCD163 in CT, CD68pCD163n overCD163 in IMCT, CD163/CD68 in IM, CD163/CD68 in CT, CD68IMCT, CD163IMCT, CD163/CD68 IN IM AND CT , CD68 0-50 TB, CD68 PER TB 0-50, CD68 0-100 TB, CD68 PER TB 0-100, CD163 0-50 TB, CD163 PER TB 0-50, CD163 0-100 TB, CD163 PER TB 0-100, CD68 CD163 WITHIN 0-50 TB, CD68 CD163 PER TB 0-50, CD68 CD163 WITHIN 0-100 TB, CD68 CD163 PER TB 0-100, CD68pCD163n 0-50 TB, CD68pCD163n 0-100 TB, CD68pCD163n PER TB 0-50, CD68pCD163n PER TB 0-100, CD163andCD68pCD163n WITHIN 0-50 TB, CD163andCD68pCD163n PER TB 0-50, CD163andCD68pCD163n WITHIN 0-100 TB, CD163andCD68pCD163n PER TB 0-100	► Number or ratio of macrophage (CD163 or CD68) or lymphocytic (CD3 or CD8) infiltrates in IM, CT, WTS or within a specified distance from TBs (in μm) CD68pCD163n: CD68 positive and CD163 negative /TB number: divided by the total number of TBs PER TB: average per each TB
DiseaseSpecificSurvival	► DSS time in months
DiseaseSpecificDeath	► Binary outcome for Disease Specific Death: 1 if died and 0 if censored
Sex	► Binary outcome for gender: 1 if Male and 2 if Female
Age	► Categorical outcome for age: 1 if less than or equal 70 years old, 2 if between 71 and 80 and 3 otherwise
pT	► Binary outcome for T Stage: 1 if pT3 and 2 if pT4
Site	► Binary outcome for location of primary tumour: 1 if left-sided, 2 if right-sided and 3 if rectal
Diff	► Binary outcome for differentiation: 1 if moderately differentiated, 2 if poor, 3 if well and other for otherwise
EMLVI	► Binary outcome for EMLVI: 1 if yes and 2 if no

Table B.2: Variables description table of spatial density and clinicopathological variables

	Train (n=113)	Validation(EDI) (n=56)	Validation(JAP) (n=61)	Overall (n=230)
DiseaseSpecificDeath				
0	98 (86.7%)	45 (80.4%)	48 (78.7%)	191 (83.0%)
1	15 (13.3%)	11 (19.6%)	13 (21.3%)	39 (17.0%)
Sex				
1	56 (49.6%)	34 (60.7%)	41 (67.2%)	131 (57.0%)
2	57 (50.4%)	22 (39.3%)	20 (32.8%)	99 (43.0%)
Age				
1	45 (39.8%)	24 (42.9%)	39 (63.9%)	108 (47.0%)
2	32 (28.3%)	14 (25.0%)	18 (29.5%)	64 (27.8%)
3	36 (31.9%)	18 (32.1%)	4 (6.6%)	58 (25.2%)
pT				
1	87 (77.0%)	46 (82.1%)	54 (88.5%)	187 (81.3%)
2	26 (23.0%)	10 (17.9%)	7 (11.5%)	43 (18.7%)
Site				
1	38 (33.6%)	9 (16.1%)	22 (36.1%)	69 (30.0%)
2	42 (37.2%)	29 (51.8%)	13 (21.3%)	84 (36.5%)
3	33 (29.2%)	18 (32.1%)	26 (42.6%)	77 (33.5%)
Diff				
1	91 (80.5%)	14 (25.0%)	20 (32.8%)	125 (54.3%)
2	19 (16.8%)	8 (14.3%)	6 (9.8%)	33 (14.3%)
3	3 (2.7%)	0 (0%)	34 (55.7%)	37 (16.1%)
Other	0 (0%)	34 (60.7%)	0 (0%)	34 (14.8%)
Missing	0 (0%)	0 (0%)	1 (1.6%)	1 (0.4%)
EMLVI				
1	18 (15.9%)	3 (5.4%)	0 (0%)	21 (9.1%)
2	95 (84.1%)	34 (60.7%)	0 (0%)	129 (56.1%)
N/A	0 (0%)	19 (33.9%)	61 (100%)	80 (34.8%)

Table B.3: Summary statistics of categorical variables. see tables B.2 and B.1 for variable coding

		Predicted	
		pT3	pT4
Actual	Censored	40	5
	Dead	6	5

Table B.4: Confusion matrix for pT stage on Edinburgh validation cohort

		Predicted	
		pT3	pT4
Actual	Censored	43	5
	Dead	11	2

Table B.5: Confusion matrix for pT stage on Japanese validation cohort

		Predicted	
		Low Risk	High Risk
Actual	Censored	46	2
	Dead	11	2

Table B.6: Confusion matrix for NB on Edinburgh validation cohort

		Predicted	
		Low Risk	High Risk
Actual	Censored	44	1
	Dead	8	3

Table B.7: Confusion matrix for NB on Japanese validation cohort

		Predicted	
		Low Risk	High Risk
Actual	Censored	42	3
	Dead	8	3

Table B.8: Confusion matrix for SVM on Edinburgh validation cohort

		Predicted	
		Low Risk	High Risk
Actual	Censored	47	1
	Dead	12	1

Table B.9: Confusion matrix for SVM on Japanese validation cohort

		Predicted	
		Low Risk	High Risk
Actual	Censored	34	11
	Dead	5	6

Table B.10: Confusion matrix for gbm on Edinburgh validation cohort

		Predicted	
		Low Risk	High Risk
Actual	Censored	40	8
	Dead	7	6

Table B.11: Confusion matrix for gbm on Japanese validation cohort

B.3 Figures

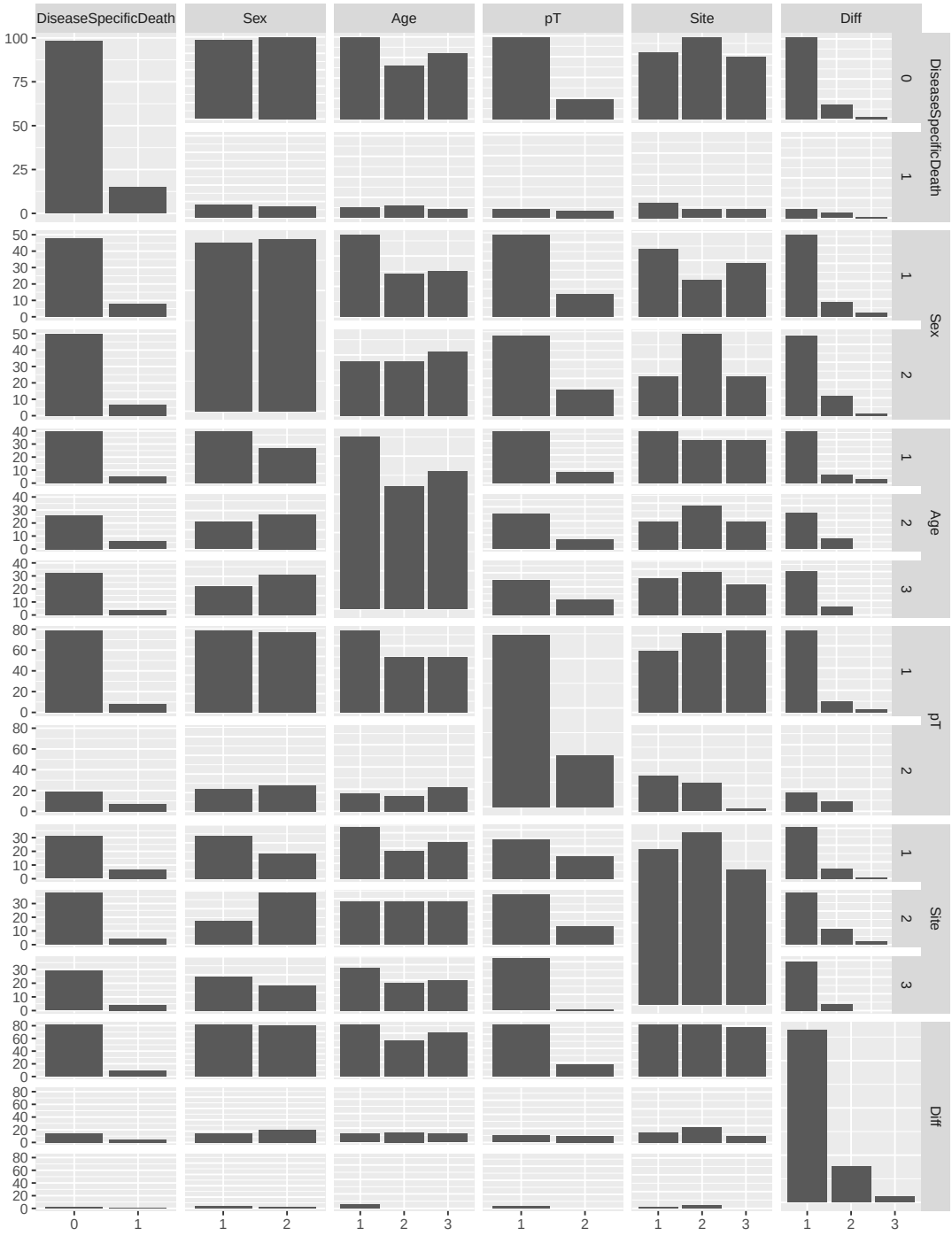


Figure B.1: Barplot matrices for categorical variables, where y axis is the frequency and x axis is the corresponding variable codes, see tables B.2 and B.1

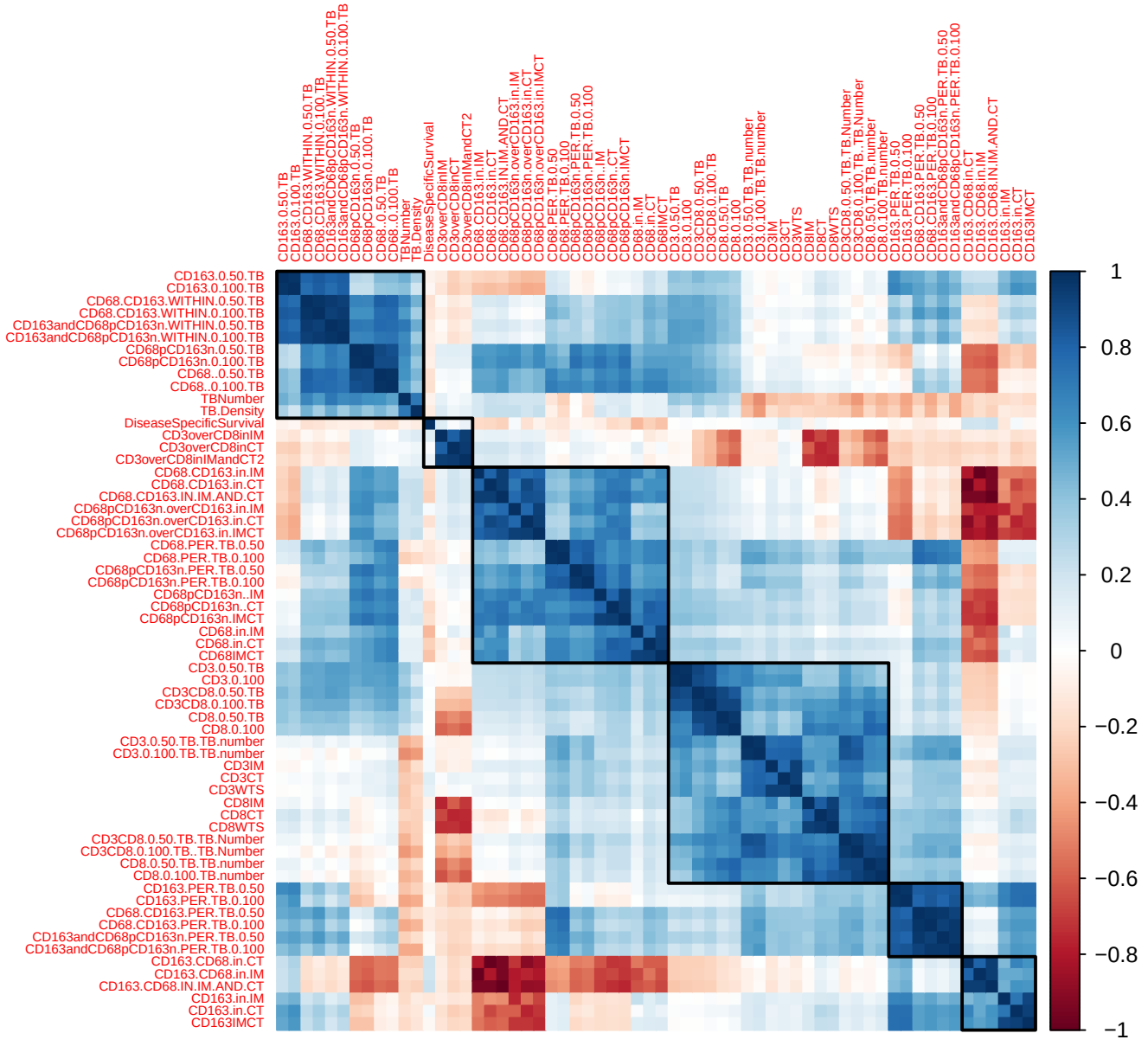


Figure B.2: Spearman rank correlation matrix between the predictors (x- and y-axis) for other predictors with Ward's hierarchical clustering. Blue colours represent high positive correlations and red colours high negative correlations (see legend). Six blocks of high positive correlation are in bold squares.

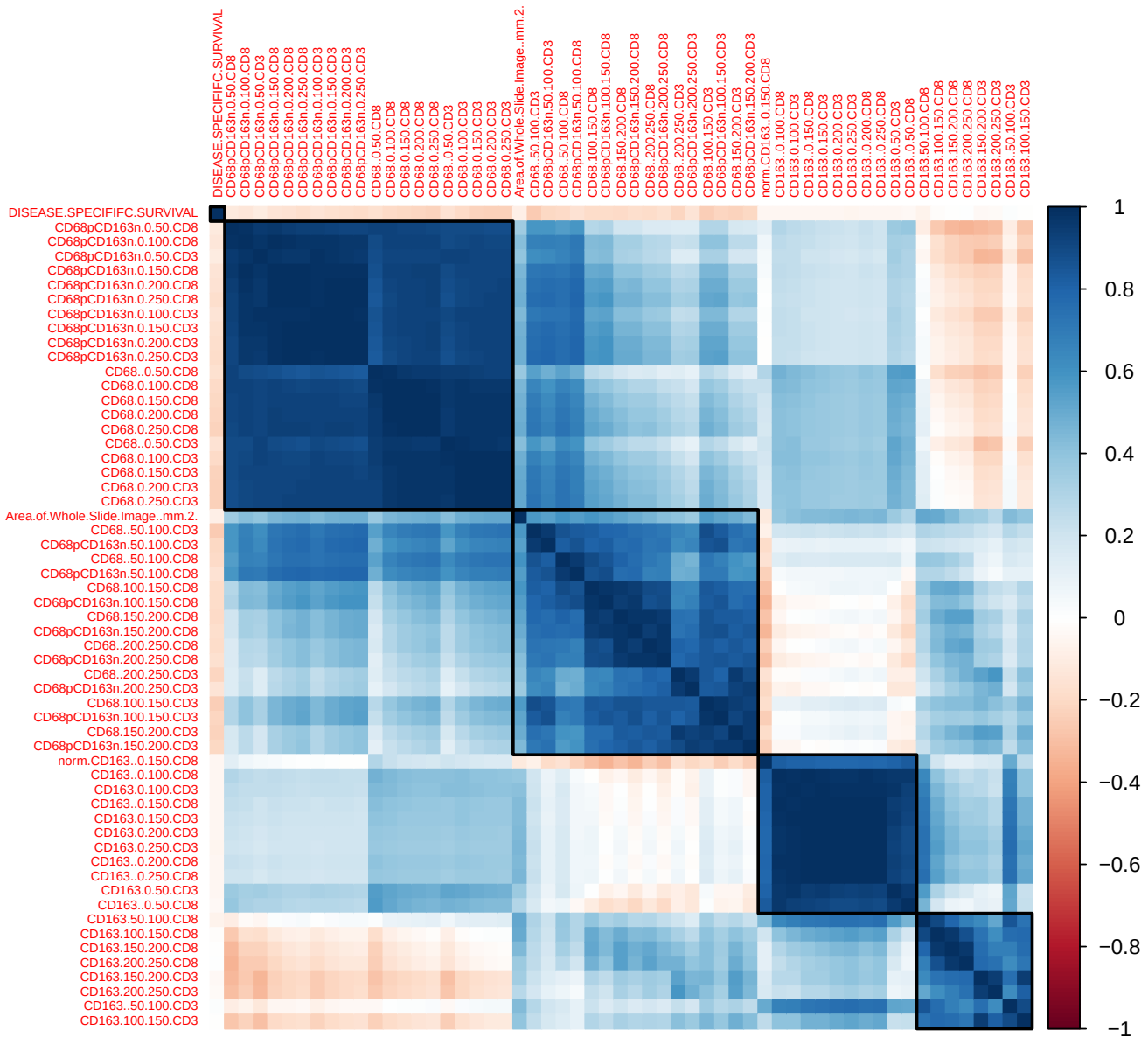


Figure B.3: Spearman rank correlation matrix between the predictors (x- and y-axis) for macrophage infiltrates with complete average linkage hierarchical clustering.
Blue colours represent high positive correlations and red colours high negative correlations (see legend). Five blocks of high positive correlation are in bold squares.

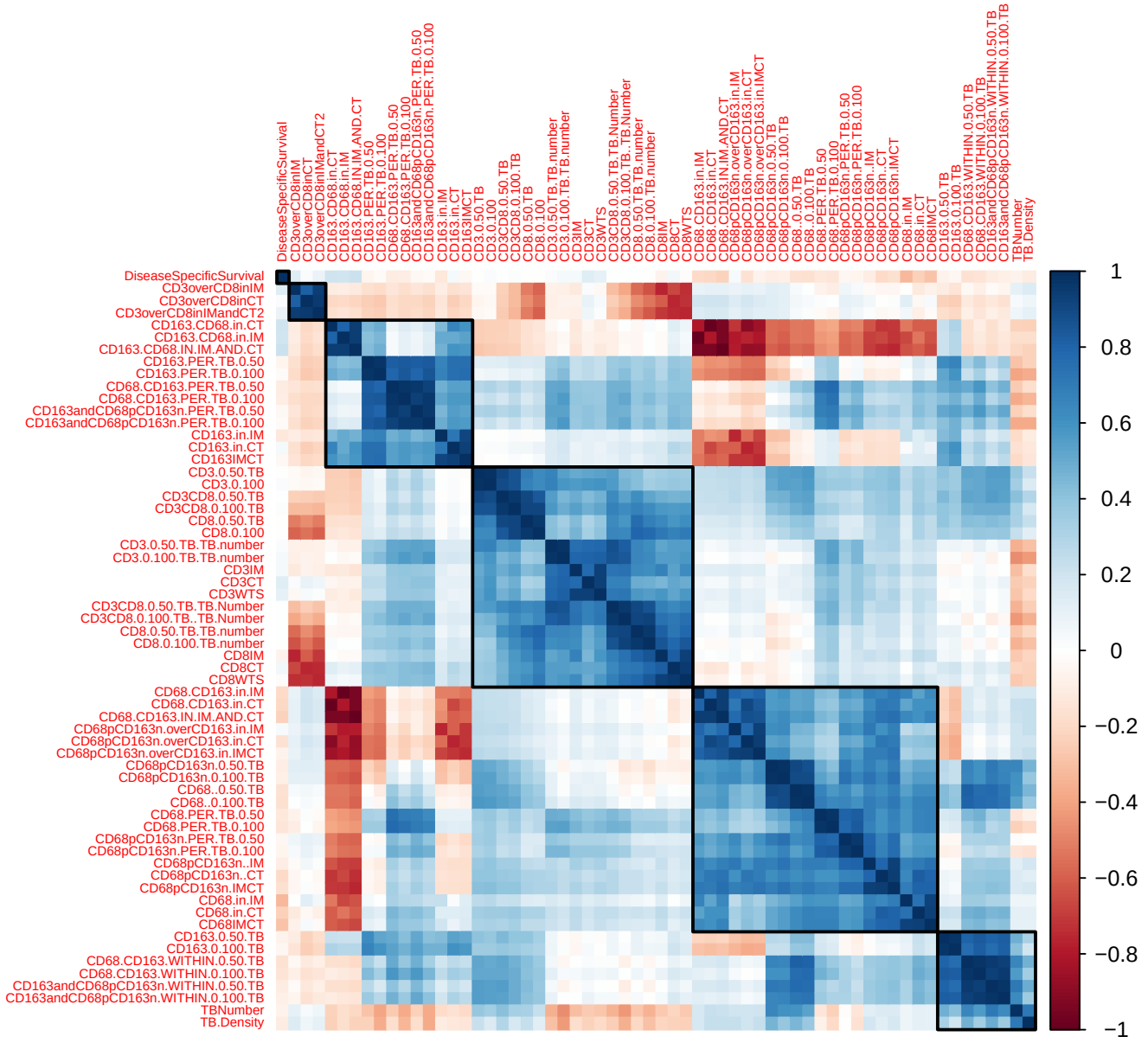


Figure B.4: Spearman rank correlation matrix between the predictors (x- and y-axis) for other predictors with complete average linkage hierarchical clustering. Blue colours represent high positive correlations and red colours high negative correlations (see legend). Six blocks of high positive correlation are in bold squares.

REFERENCES

[1] Cancer Research UK. Bowel cancer statistics | Cancer Research UK; 2019. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer{#}heading-Zero.

[2] Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. The Lancet. 2018;391(10135):2128–2139.

[3] Nearchou IP, Lillard K, Gavriel CG, Ueno H, Harrison DJ, Caie PD. Automated Analysis of Lymphocytic Infiltration, Tumor Budding, and Their Spatial Relationship Improves Prognostic Accuracy in Colorectal Cancer. Cancer immunology research. 2019;7(4):609–620.

[4] Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine. 1997;16(4):385–395.

[5] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. vol. 27. Springer; 2005.

[6] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301–320.

[7] Tibshirani R, Wainwright M, Hastie T. Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC; 2015.

[8] Zou H. The adaptive lasso and its oracle properties. Journal of the American statistical association. 2006;101(476):1418–1429.

[9] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67.

[10] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267–288.

[11] Chen S, Donoho D. Basis pursuit. In: Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers. vol. 1. IEEE; 1994. p. 41–44.

[12] Park T, Casella G. The bayesian lasso. Journal of the American Statistical Association. 2008;103(482):681–686.

[13] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association. 2001;96(456):1348–1360.

[14] Meinshausen N, Bühlmann P. Variable selection and high-dimensional graphs with the lasso. Annals of Statistics. 2006;34:1436–1462.

[15] Lee JD, Sun DL, Sun Y, Taylor JE, et al. Exact post-selection inference, with application to the lasso. The Annals of Statistics. 2016;44(3):907–927.

[16] Taylor J, Tibshirani R. Post-selection inference for-penalized likelihood models. Canadian Journal of Statistics. 2018;46(1):41–61.

[17] Tibshirani RJ, et al. The lasso problem and uniqueness. Electronic Journal of Statistics. 2013;7:1456–1490.

[18] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33(1):1.

[19] Huang J, Ma S, Zhang CH. Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica. 2008;p. 1603–1618.

[20] Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Annals of statistics. 2009;37(4):1733.

[21] Meinshausen N. Relaxed lasso. Computational Statistics & Data Analysis. 2007;52(1):374–393.

[22] Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187–202.

[23] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for CoxâĂŹs proportional hazards model via coordinate descent. Journal of statistical software. 2011;39(5):1.

[24] Breiman L. Random forests. Machine learning. 2001;45(1):5–32.

[25] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS, et al. Random survival forests. The annals of applied statistics. 2008;2(3):841–860.

[26] Segal MR. Regression trees for censored data. Biometrics. 1988;p. 35–47.

[27] LeBlanc M, Crowley J. Survival trees by goodness of split. Journal of the American Statistical Association. 1993;88(422):457–467.

[28] Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015.

[29] Ishwaran H, Kogalur U. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package version; 2019.

[30] Greenwell B, Boehmke B, Cunningham J, Developers G. gbmâĂŹ: Generalized Boosted Regression Models. R Package Version 21. 2018;4.

[31] Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001;p. 1189–1232.

[32] Mason L, Baxter J, Bartlett PL, Frean MR. Boosting algorithms as gradient descent. In: Advances in neural information processing systems; 2000. p. 512–518.

[33] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. vol. 112. Springer; 2013.

[34] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–357.

[35] Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien; 2019. R package version 1.7-0.1. Available from: `https://CRAN.R-project.org/package=e1071`.

[36] from Jed Wing MKC, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al.. caret: Classification and Regression Training; 2018. R package version 6.0-81. Available from: `https://CRAN.R-project.org/package=caret`.

[37] Jones-Todd CM, Caie P, Illian JB, Stevenson BC, Savage A, Harrison DJ, et al. Identifying prognostic structural features in tissue sections of colon cancer patients using point pattern analysis. Statistics in medicine. 2019;38(8):1421–1441.

[38] Torgo L. Data Mining with R, learning with case studies Chapman and Hall/CRC. URL: http://www dcc fc up pt/˜ ltorgo/DataMiningWithR. 2010;.

[39] Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. R journal. 2014;6(1).

[40] Ehrlinger J. ggRandomForests: Visually Exploring Random Forests; 2016. R package version 2.0.1. Available from: `https://CRAN.R-project.org/package=ggRandomForests`.

[41] Harrell Jr FE, with contributions from Charles Dupont, many others. Hmisc: Harrell Miscellaneous; 2019. R package version 4.2-0. Available from: `https://CRAN.R-project.org/package=Hmisc`.

[42] Liaw A, Wiener M, et al. Classification and regression by randomForest. R news. 2002;2(3):18–22.

[43] Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab-an S4 package for kernel methods in R. Journal of statistical software. 2004;11(9):1–20.

[44] Majka M. naivebayes: High Performance Implementation of the Naive Bayes Algorithm; 2019. R package version 0.9.5. Available from: `https://CRAN.R-project.org/package=naivebayes`.

[45] Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. Springer Science & Business Media; 2013.

[46] Therneau T. A Package for Survival Analysis in S. version 2.38; 2015.

[47] van Wieringen WN. Lecture notes on ridge regression. arXiv preprint arXiv:150909169. 2015;.