Regularized Risk Prediction Models in Subject/Patient Analytics in a Time to Event Setting

Wilbur Zhu

A thesis submitted in partial fulfilment
of the requirements for PhD
of University College of London

Department of Computer Science
University College London

2021

Statement

I, Wilbur Zhu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I declare that this has been indicated on the thesis.

Acknowledgments

I would not have been able to be in this position today without so many for whom I am forever grateful for.

My supervisors Professor Philip Treleaven and Dr Nick Firoozye for their supervision, especially during the pandemic.

My parents for their sacrifices throughout the years for my education and their support.

My examiners Professor Philippe de Wilde and Professor David Ingram for their thorough and fair examination that has improved this thesis further.

My many friends- from those who encouraged me to never give up on my academic dreams to those whom I serve alongside in the British military who helped me overcome my fears.

Finally, I would like to thank this great country for giving both me and my parents the opportunities we've had. Its values and people are what motivate me to dedicate the rest of my life to service to this country.

Abstract

This thesis comprises of five investigations and focuses on the use of risk prediction modelling from a computational statistics and machine learning perspective, with applications in subject (e.g. gym user, patient) analytics in a time to event setting. The work was conducted in collaboration with eGym and UCL Hospitals (UCLH). A variety of computational statistics (e.g. logistic lasso) and machine learning based risk prediction methods are applied ranging from kernel methods, ensemble methods and decision trees from both a classification and survival perspective. The thesis is concerned with modelling gym user behaviour and predicting treatment times and types. The underlying goal of this thesis is to develop generalizable and useful models to predict gym user behaviour and patient treatment times. This is what leads us to our methodological work in chapter 6.

This thesis conducts the following investigations.

1. **Weibull full likelihood implementation**

   The first investigation involves conducting an implementation of a Weibull full likelihood survival model in R. The aim of this investigation is to build the Weibull distribution proportional hazards model, which is formulated via the log likelihood. Then we apply this model to simulated data to see whether the model can reveal the real pattern of the data. The results prove that from the synthetic data the model we build in R can unearth the parameters and the coefficients from which we generate the data.

2. **Predicting gym user behaviour through churn and visits**

   The second investigation consisting of two sub-investigations considers the use of time to event models to predict gym user behaviour and churn. The data set has been provided by the Gym Equipment manufacturer eGym. The first sub-investigation considers if it is possible, we can predict whether or not a user will churn, using a range of methods across computational statistics and machine learning, from logistic regression to survival random forests. Our findings indicate that with demographics alone we are unable to produce machine learning models that outperform a baseline learner. This tells us that we are unable to predict right at the beginning, whether or not a user will churn. However, when we apply machine learning based survival models including elastic net Cox and Cox Boosting, we are able to outperform the baseline. This sub-investigation serves as an introduction to considering gym user churn in a time to event setting through both classification and survival models. In the second sub-investigation, we then apply risk prediction modelling in predicting gym user visits via a moving window model, we find we are marginally able to outperform the majority vote baseline in some settings.

3. **Predicting patient treatment times and treatment types for patient rehabilitative care**

   The third investigation, also consisting of two sub-investigations, concerns the use of time to event modelling to predict patient treatment times and treatment types for patient rehabilitative care. The underlying goal is to help design treatment plans aimed at helping patients return to work by predicting the required combination of treatment time and treatment types required for each patient. The data has been provided by UCLH. All patients in the data set have been eventually discharged from the treatment programme. The aim of the first sub-investigation is to predict how much treatment time the patients required before they were discharged and which patients are more likely to take longer. We model this problem using regression and survival analysis, methods used range from generalized additive models to Cox boosting. Our results show that, using demographic variables we are able to outperform the baseline. In the second sub-investigation, we utilise risk prediction models, such as logistic regression and Adaboosting to predict treatment types based on demographics. We are able to outperform the baseline for some treatments in a deterministic setting but not in a probabilistic setting.

4. **Regularization problems in gym user/patient setting**

   As alluded, in both our application settings our model performances are mixed. Our aim therefore is to investigate how we can potentially improve our model performance and usefulness. This is what motivates our methodological studies: improving our model performance via hyper-parameter tuning based on the relevant loss function. We begin our investigation by using F1, Brier score and net benefit as the scoring functions for parameter tuning to build LASSO models. We then run the models on the gym user data and hospital data and compare the performance outputs from modelling. We find we are able to outperform the conventional LASSO models in terms of F1, Brier score and net benefit when using them as tuning functions, respectively. The different LASSO models provide different variable selections and insights. Then we use the integrated Brier score to turn the parameters of Cox proportional hazards LASSO models in a survival setting. Compared with the conventional performance measure - Concordance index, the integrated Brier score reflects better the error measure overall time. We find that by tuning parameters for the integrated Brier score we are able to obtain better integrated Brier score performance and different variable selections. We also examine whether the integrated Brier score is not only useful for improving survival performance at all times but at specific times too. We apply the Cox proportional hazards LASSO models with integrated Brier score and Concordance index as the scoring functions to the gym user and hospital data sets. The results show the models can better perform on the corresponding loss functions but the integrated Brier score LASSO model doesn't guarantee better performance at a specific time. Finally, we extend our methodology to more modern machine learning methods such as support vector machines. We use F1 score, Brier score and net benefit as scoring functions to turn the parameters γ and C

of SVMs and run the models on the gym user data and hospital data. The results show they only slightly outperform the conventional model and are specifically poor in the deterministic setting due to the data imbalance.

Contributions to Science

This thesis makes the following contributions to science.
1. Applies logistic regression, linear discriminant analysis, support vector machines and random forests to predict the gym user attendance and churn.
2. Introduces the idea of comparing gym user prediction models to a majority vote baseline.
3. Introduces moving window prediction models for gym user visit prediction.
4. Discovers the relationship between patient demographics and rehabilitative care treatment times.
5. Introduces machine learning and computational statistics to predict patient treatment times and types for neurological rehabilitation patients.
6. Introduces the use of the F1, Brier score and in particular the net benefit LASSO models to a gym user churn prediction and a treatment type prediction.
7. Introduces the use of the integrated Brier score for tuning Cox LASSO models.
8. Extends the idea of parameter turning via the F1, Brier score and net benefit to modern machine learning methods.

Impact Statement

Analytics is transforming wellbeing and healthcare. The context for impact of time to event models in a health and wellbeing setting is driven by the explosion in data collection capabilities. This can range from data collected by wearable technologies (Dinh-Le et al., 2019) and electronic health records (Kalra & Ingram, 2006). Often such data sets enable us to extract events of interesting in healthcare such as a time to relapse of cancer (Zeng et al., 2019) or the injury of an athlete (Zadeh et al., 2020) whilst also providing predictive variables for the models. These can be seen as time to event problems. The thesis addresses both healthcare as well as methodological impacts. From an applied perspective this thesis contributes gym user analytics and rehabilitative patient treatment prediction. The gym user analysis introduces a wide variety of machine learning based strategies that can be applied not only to gym user behaviour but any exercise app that records attendance. Our second application impact is to introduce machine learning to modelling rehabilitative care for neurological patients. From a methodological perspective we introduce two new tools for improving the utility of risk prediction models. The net benefit LASSO enables us to build a risk prediction model that weighs any potential costs of false positives or false negatives. The integrated Brier score LASSO allows us to tune survival LASSO models over time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides an introduction to the research topic. We start by giving a motivation to both the research area and the applications. We then proceed to motivate the research by describing the two data sets which this thesis is based on.

This thesis examines the application of machine learning based risk prediction modelling to patient analytics.

1.1 Research Motivation

The context for studying time to event models in a health and wellbeing setting is driven by the explosion in data collection capabilities. This can range from data collected by wearable technologies (Dinh-Le et al., 2019) and electronic health records (Kalra & Ingram, 2006). Often such data sets enable us to extract events of interest in healthcare such as a time to relapse of cancer (Zeng et al., 2019) or the injury of an athlete (Zadeh et al., 2020) whilst also providing predictive variables for the models.

The motivations for this thesis are both methodological and applied. From an applied perspective we are seeking to use risk prediction to improve health and well-being firstly through analysing exercise behaviour when a user uses a gym and secondly creating effective treatment plans to help patients get back to work.

The first problem we are interested in is to encourage an active lifestyle. To achieve this, we

must be able to understand and predict gym user behaviour. This involves predicting both whether or not a gym user will stop attending the gym completely - churn, and also whether or not a gym user will attend a gym in a given time.

The second problem of interest stems from helping patients cope with long-term health conditions and return to a better-quality lifestyle. Neurological conditions often have long-term effects and therefore patients not only need treatment for their condition but also treatments to help them return to work. Such treatment programmes often require a wide range of different types of treatments and rely on a wide range of factors.

Both problems provide motivation for the methodological research in this thesis. We want to develop methods that are generalizable and can be reproduced on future data set. In the modelling process, we use the existing data to build our models. The data not only contains information representing the real data pattern and also contains a certain noise. Therefore, we aim to ensure that our models only fit the real data pattern with the minimum effect of noise. In addition, we want to build models that are interpretable and provide better insights into the data. This is achieved through our methodological work on regularized models.

This thesis addresses risk modelling motivated by two data sets. The first data set is in the area of user behaviour and concerns gym user attendance and therefore the event of interest will be whether or not a gym user attends the gym. The second data set lies in the healthcare domain and concerns a rehabilitative care treatment programme and the event of interest is the point at which the patient is discharged. This thesis comprises five investigations and focuses on the use of risk prediction modelling from a computational statistics and machine learning perspective, with applications in user/patient analytics in a time to event setting. The work is conducted in collaboration with eGym and UCL Hospital (UCLH). A variety of computational statistics (e.g. logistic lasso) and machine learning based risk prediction methods are applied, ranging from kernel methods, ensemble methods and decision trees from both a classification and survival perspective. The thesis is concerned with modelling gym user behaviour and predicting treatment times and types. The underlying goal of this thesis is to develop generalizable and useful models to predict gym user behaviour and patient treatment times. This is what leads us to our methodological work in chapter 6.

1.1.1 Motivating Data Set and Applications 1: eGym Data Set

The data set is provided by a German Gym equipment manufacturer - eGym. When a gym user uses their machines or mobile app first time, the user must create an account that collects the user's demographic information. Thereafter every time they use the machines they must sign

in. This produces a data set containing over 1 million users with their demographic information and a time series of gym sign ins. The event of interest is therefore whether or not a user will visit the gym. This motivates us to develop models to predict gym user behaviour. The data set contains the following variables:

Demographic Variables: age, gender, gym location, subscription level,

Attendance: time series for sign ins.

1.1.2 Motivating Data Set and Applications 2: UCL Hospital (UCLH) Data Set

This data set is provided by the UCLH neurological ward who have implemented a neurological rehabilitation programme. The events of interest are to whether or not patients require a certain treatment and the time at which a patient is discharged. The data set contains 442 users with the following variables:

Demographic Variables: age, gender, ethnicity, marital status, education level,

Work-Related Variables: occupation type, pre-injury work status, pre-injury work hours, initial work hours, initial work status,

Questionnaire Responses: self-care, mobility, pain, usual activities, anxiety,

Treatment Times: total treatment time, time for face to face with psychologist, time for face to face with occupational health therapist.

The aim is to use patients' demographic and work-related information to predict how long the treatment time and what type of treatment a patient requires.

1.2 Research Objectives

This section details what we hope to achieve through our investigations.

1. Our first objective is to examine how we can optimize a full log likelihood proportional hazards model using a R library with an assumed distribution. The distribution we chose is the Weibull distribution for its versatility and simplicity. Our aim is to examine the 'quality' of the implemented model by testing how well the log likelihood optimization is able to fit the Weibull simulated data. This will enable us to examine the feasibility of using this function to conduct our proposed future work of optimized penalized likelihood functions.

2. Our second objective is to investigate the feasibility of applying risk prediction models to predict gym user behaviour. The goal is to see whether or not risk prediction models can outperform a simple 'majority vote' model in predicting gym user behaviour.

3. Our third objective is to examine whether or not we can use machine learning methods to help build better treatment plans for rehabilitative care. The goal is to examine whether or not risk prediction models can help predict how much treatment time a patient requires and whether or not a specific treatment is required.

4. Our last objective is to investigate whether we can improve the gym user churn prediction and patient treatment type prediction by using F1, Brier score and net benefit LASSO.

1.3 Research Investigations

This section provides a summary of the investigations conducted along with our findings.

1.3.1 Weibull Full Likelihood Implementation

The first investigation involves implementing a Weibull full likelihood proportional hazards survival model in R. As the full log likelihood in the Weibull proportional hazards model is differentiable, we use Quasi-Newton method (also known as a variable metric algorithm) as the optimization algorithm. The performance of the implementation is studied on Weibull simulated data. We first test the performance on a simulated data set of a fixed sample size. We find that the model is able to fit well onto the simulated data. We then simulate the data sets on varying sample sizes and look at the mean squared error. We find that the mean squared error and its standard deviation decrease as sample size increases. This is consistent with the law of large numbers and the properties of variances. The results demonstrate that the Quasi-Newton algorithm is a viable option for fitting differentiable likelihood functions.

1.3.2 Predicting Gym User Behaviour through Churn and Visits

The second investigation, which consists of two sub-investigations, considers the use of time to event models to predict gym user behaviour and churn. The data set has been provided by a German gym equipment manufacturer - eGym.

1. **Predicting gym user churn:** This investigation considers whether or not we can predict if a user will churn, i.e. the user stops attending the gym completely. We predict churn from a classification and survival perspective. In the classification setting we use logistic regression, linear discriminant analysis, random forests and support vector machines. We find that based on demographics information in the data set, we are unable to produce machine learning models which outperform a baseline learner, i.e. the majority vote. This tells us that we are unable to predict right at time when a user signs up, whether or not a user will churn. Even when we applied the same models with visit information as an additional variable, we are still unable to outperform a baseline learner. However, in the survival setting, when we apply machine learning based survival models including elastic net Cox and Cox Boosting we are able to outperform the baseline learner.
2. **Predicting gym user behaviour:** In this investigation we then apply risk prediction models in the context of predicting gym user visits. By applying the same classification methods used in the churn prediction setting, we find we are able to outperform the majority vote baseline in some models in terms of accuracy, log loss and Brier scores but not others.

1.3.3 Predicting Patient Treatment Times and Treatment Types for Patient Rehabilitative Care

The third investigation involves the use of computational statistics and machine learning based risk prediction models to help create treatment plans. The investigation introduces machine learning models to patient rehabilitative care in a neurological setting. Again, this investigation consists of two sub-investigations.

1. **Modelling treatment times:** The aim is to determine which demographic factors help determine treatment times. We achieve this by applying stepwise selection and LASSO models from both a survival and regression modelling perspectives. Our next step is to see if we could apply survival and regression models to predict treatment times. We apply generalized linear models, elastic net models, generalized additive models and compare

the performance against a featureless baseline model (model without any covariates and just an intercept). Our models are able to perform better against a baseline model. In the survival setting we apply Cox boost, Cox proportional hazards and the elastic net Cox. Our models are able to outperform the baseline model.

2. **Predicting treatment types:** In this investigation we apply risk prediction models to predict treatment types based on demographics. We use logistic regression, linear discriminant analysis, random forests and support vector machines to predict whether or not a patient will require a specific treatment type. We find that our models are unable to identify whether or not the patient requires a specific treatment.

1.3.4 F1, Brier Score, Net Benefit and Integrated Brier Score for Parameter Turning

Having tried to predict gym user churn and patient treatment, we explore ways in which we can improve our prediction models and gain better insights by using more interpretable models. This investigation comprises three sub-investigations.

1. **F1, Brier score and net benefit LASSO logistic models:** The aim of this investigation is to investigate the properties of alternative tuning strategies for logistic LASSO models using different loss functions. The goal is to compare their performance and properties to determine whether there are any potential benefits of such approaches. We run these models on the gym user and hospital data sets and find that using the F1, Brier scores and net benefit for parameter turning in LASSO modelling we are able to obtain better performances and alternative interpretations.

2. **Integrated Brier score Cox LASSO model:** The purpose of this investigation is to introduce the integrated Brier score into Cox LASSO modelling to see whether we can improve the model performance. We build the Cox LASSO model with integrated Brier score as the scoring function to turn the parameter λ and apply the models to the gym user and hospital data sets. The results show the integrated Brier score LASSO Cox model is able to outperform the conventional Concordance index LASSO Cox model.

3. **F1, Brier score and net benefit SVM models:** In this investigation, we extend our exploration to more modern machine learning models. Using support vector machines as an example, we use F1, Brier score and net benefit as the scoring function to turn parameters γ and C for SVM models. We apply the models to the gym user and hospital data sets and the results are mixed and even poor in deterministic measures. This is because the

data sets we use are severely class imbalanced. In addition, the performance of the models is likely to have been impacted by the quality of data - lack of explanatory variables that can fully explain the observed phenomena and large amounts of missingness in the data.

1.4 Methodological Questions to Address

1. **Performance measure optimization through parameter tuning:** The first methodological question to answer is to investigate whether we can optimize a performance measure by tuning using that performance measure. We focus on two classification methods - logistic LASSO and support vector machines. For the logistic LASSO the central question is - can the logistic LASSO performance be improved for a given loss function if we tune using that loss function. The loss functions of interest are the F1, Brier score and net benefit. We then aim to extend this question to the parameter tuning of support vector machines (SVMs).

2. **Censored performance measures in survival LASSO models:** The second methodological question is to investigate how the performances of survival LASSO models tuned via performance measures that account for censoring perform compared to performance measures that do not account for censoring, on heavily censored data. The aim is to focus on the Cox proportional hazards LASSO and the use of two loss functions the Concordance index (which cannot account for censoring) and the Integrated Brier Score (which accounts for censoring).

3. **Possible extensions:**

   Both methods could be extended by considering alternative models and loss functions. For the first methodological question the most obvious extension would be to use ridge or elastic net models in addition to the LASSO models. For the second methodological question, alternative models such as full likelihood proportional hazards models and accelerated failure time models can be used. These models would have the additional benefit of being computationally easier. Alternative loss functions include the integrated log loss or the Kent and Quigley R squared measures.

1.5 Contributions to Science

This section summarizes the contributions to science.

1. Methodologically this thesis investigates the use of the optimization function available in R to help fit full likelihood survival models. Using the Weibull survival model as an example we find that the full likelihood Weibull model we build from the log likelihood equation is able to uncover the underlying patterns from a simulated data.

2. This thesis investigates the viability of applying machine learning risk prediction models to predict and understand gym user behaviour. This consists of two parts. Firstly, we examine the feasibility of predicting if a user will stop visiting the gym. The second part of this investigation aims to understand how we can predict whether or not a gym user will visit. Again, we find that outperforming a majority vote baseline is difficult. We find we are able to outperform the majority vote baseline in some settings but not others.

3. In our second application setting we are able to develop models to help create better treatment plans. We find that we are able to use feature selection to find the relevant variables that help determine treatment times. We are also able to create models that can outperform an intercept only model. We then aim to predict different treatment types. We found that since for each treatment the majority did not require it, predicting whether or not a patient would require a specific treatment is difficult, even with machine learning methods.

4. This thesis introduces the use of the F1, Brier score and net benefit LASSO in a clinical prediction setting.

5. This thesis extends the idea of optimizing for the net benefit for other machine learning models.

6. The research work in this thesis contributes to the open source library by providing an implementation of Weibull full likelihood proportional hazard model, F1, Brier score and net benefit LASSO models, integrated Brier score LASSO Cox proportional model and F1, Brier score and net benefit support vector machine models.

1.6 Thesis Structure

Chapter 2 Research Questions and Methodology Review begins by describing the risk prediction framework from both a qualitative and quantitative perspective. We also provide an introduction to the concept of overfitting and regularization to provide the motivation for my future work. We then proceed to give a literature review in risk prediction in the various settings across a variety of domains. Finally, we end by giving a review into the different ways to evaluate the performance of risk prediction models.

Chapter 3 Weibull Full Likelihood Implementation provides an in depth review of survival models - one of the various risk prediction approaches introduced in chapter 2. We describe the various types of survival models and the various ways in which they can be fitted. We then proceed to implement a Weibull survival model using the inbuilt optimizer in R. We evaluate the performance of our model using simulated data.

Chapter 4 Predicting Gym User Behaviour through Churn and Visits describes the investigation of section 1.3.2 Predicting Gym User Behaviour through Churn and Visits. The chapter begins by giving an overview of the problem. We then proceed to data cleaning and data processing. Afterwards we carry out exploratory analysis of the cleaned data sets. Finally, we conduct predictive modelling, using logistic regression, linear discriminant analysis, random forests and support vector machines.

Chapter 5 Predicting Patient Treatment Times and Treatment Types for Patient Rehabilitative Care details the investigation of section 1.3.3 Predicting Patient Treatment Times and Treatment Types for Patient Rehabilitative Care. The chapter begins by describing the research question. We then proceed to conduct variable selection to find the variables that impact treatment times. Finally, predictive models are applied to the data set from both a regression and survival modelling point of view.

Chapter 6 Regularization Problems in a Gym User/Patient Setting narrates the investigation of section 1.3.4 F1, Brier Score, Net Benefit and Integrated Brier Score for Parameter Turning. We start by discussing potential benefits of using the F1, Brier score and net benefit as scoring functions in logistic LASSO modelling. Next, we introduce the integrated Brier score into the Cox proportional hazard LASSO model for parameter turning. Finally, we extend the F1, Brier score and net benefit for parameter turning to support vector machines. We build these models and apply them to the gym user and hospital data sets and analyse the results.

Chapter 7 Conclusion and Future Work details Methodological Questions to Address.

Chapter 2

Research Questions and Methodology Review

This chapter provides an overview of the risk modelling framework. We begin by giving a qualitative overview of risk modelling. Next, we will give a mathematical formulation of the three main approaches to risk prediction. We will then proceed to review the literature surrounding risk prediction models. Finally, we will give an overview of the major performance evaluation methods in risk prediction.

2.1 The Risk Modelling Framework

In this section we will give the descriptions and motivations of the risk modelling approach. We start describing the data sets. A typical data set in a risk modelling study contains observations from a number of units where the primary interest is in making predictions about a particular event - often in a practical setting this is either an adverse event or a desirable event which one wants to avoid or pursue. An example of an event of interest could be a patient being diagnosed with a disease or a machine breaking down. This event of interest is usually measured over a predetermined length of time for which the data are observed. In addition to the event status, each unit may have observed characteristics that are measured which may or may not be related to the risk of the event of interest. Over the predetermined time interval, the characteristics could be static, for example the genetic markers associated with a disease, or time varying, such as the blood pressure readings of a patient taken at various time intervals.

The main aim of risk modelling is to make predictions about whether the event will occur in the predetermined time period (and potentially how many times), the probability of the event happening, or the time until the event happens. From a methodological point of view, the first two types of prediction can be classed as deterministic and probabilistic classification problems respectively, whereas the latter is an example of survival analysis.

2.2 Approaches

In this section we begin by describing mathematically the data setting. Next, we give a motivation behind the supervised learning approach to risk prediction, then we proceed to outline supervised learning and finally we give the mathematical formulation of the three main approaches to risk prediction that we outlined previously, namely deterministic classification, probabilistic classification and survival modelling. The notation in this section is inspired by (Hastie et al., 2009).

Consider a data set which contains N units, for each individual unit i we observe p characteristics that can be observed either statically or over time. The measurements for the characteristics for each individual i are stored in a vector of length p known as x_i, x_i can contain static and non-static variables. We refer to this vector of characteristics as our covariates/variables. We also observe the individual i event status δ_i $\{0, 1\}$, where 1 is if the event happens, 0 otherwise in a predetermined time interval. Finally, we observe a time to event - T_i if the event happens for i. On the other hand, if the event doesn't happen in the predetermined length of time we define unit i to have been *censored* and we observe a time to censoring C_i for i. The time to censoring is the time from which the unit i was first observed to the end of the observation. Every unit will therefore at least have a t_i which we define as the time to event or censoring, whichever happens first. Depending on the risk prediction task in question, the event status or the event status plus time to event or censoring will be our y_i. The purpose of risk prediction modelling aims to predict the risk of such an event so that in our domain of interest, the risk for an adverse event can be mitigated or the probability for a desirable event can be strengthened.

As mentioned before, there are multiple approaches to how we predict risks. In addition to the multiple approaches, for each approach there are a wide variety of methods we can use. Irrespective of which method and approach we use, building a risk prediction model involves finding a function that maps the relationship between the input variables x_i and the event status (and time to event if applicable) y_i. Given the multiple methods for each approach, we also need to be able to choose the best method for the risk prediction problem in question. When choosing which specific method to use, we want to choose the method that has the best

performance on unseen data. The importance of evaluating our methods on unseen data is to ensure that we choose the method that is the most generalizable so that in the future we can apply this method to help mitigate the adverse event or ensure the desirable event in question. This leads to the supervised learning framework.

2.2.1 Building a Supervised Learning Task

For a risk prediction problem on a small dataset we are only able to determine the relationship between the covariates and the event of interest. However, when having larger data-sets with more observations available we are not only able to determine the relationship but also verify whether such relationship exists on new data. In addition, a data set with a large number of observations/examples also allows to potentially use alternative methods to discover more complex relationship between the covariates and event of interest. Both these factors provide motivation to consider risk prediction as a supervised learning task. The idea of supervised learning is to estimate a function that maps from the inputs (in our case the covariates) onto an output (in our case the event status). As detailed previously, we want to find a function that is the most generalizable and enables us to apply this function on future new data sets, i.e. unseen data. Therefore, we are interested in estimating the generalization or test error - the performance of the mapping function on unseen data. This is defined mathematically later on in this chapter. Let us formulate mathematically the data setting for our prediction task. Consider a set of N units with $\langle(x_1, y_1), \ldots (x_N, y_N)\rangle$, where x_i is the vector of the covariates of unit i and y_i are the target variables of unit i. The definition of y_i is dependent on the approach in question. As defined in the approaches section y_i is the event status or the event status plus time to event or censoring. In addition, we let $x_i \in X$ and $y_i \in Y$ where X, Y are our input and output spaces respectively.

As detailed in the earlier subsection we need to find appropriate function mappings from X to Y and evaluate their performance on unseen data. Therefore, in a given data set of N units we need to save a part of our data to be unseen whilst allowing the rest of the data set to be used to derive the appropriate function mapping. The set of units from which we derive the appropriate function is known as the 'training' set. Let this set be of size m, the remaining $N - m$ units which are used to evaluate the performance of each method is known as the 'test' set. We will use index i to refer a unit in the data set and use j or k to refer a unit in the training set or the test set respectively. For any method we use, irrespective of the task, the supervised learning process is to learn a suitable mapping functions to map input X to our output space Y. The mapping function can be one that directly maps X onto Y or a function that gives the probability of obtaining the output in Y. Let G be the set of such functions g that are able to do such a mapping where

$$g : X \rightarrow Y. \tag{2.1}$$

We have $\hat{g} \in G$, mapping x_i to $\hat{y_i}$. Given the outcomes Y and $\hat{g}$ we can calculate a loss function. A loss function L is a measure of how good a prediction model $\hat{g}$ is in terms of mapping input X to output Y. A mathematical formulation of a loss function is formulated in section 2.4.2 Performance Measures and Loss Functions. Using the loss function L we are able to calculate the test error,

$$\mathsf{TestError} = \frac{1}{N-m} \sum_{k=1}^{N-m} L(\hat{y_k}, y_k). \tag{2.2}$$

This is the average of the loss function over unseen data i.e. a test set. We choose the function with the lowest test error. The nature of the function g and the loss function L are dependent on the risk prediction approach chosen.

2.2.1.1 Deterministic and Probabilistic Prediction

In the deterministic setting, for each individual unit j in the training set, we attempt to predict an appropriate event status $\hat{y}_j$ given the variables x_j. Let our g take the form f be a mapping function from a family to map $X \rightarrow Y$ where X is our input covariates and Y is our event status. This can be seen as statistical classification as introduced by (Fisher, 1936). The deterministic classifier therefore takes the form,

$$\hat{y}_j = \hat{f}(x_j), \tag{2.3}$$

where $\hat{f}$ is chosen from a family of functions f from a specific deterministic classification method. The way $\hat{f}(x_j)$ is fitted on the training set is dependent on the method used.

On the other hand, when we want a risk of event to be predicted, the probabilistic classification is appropriate where our g is a probability. Probabilistic classification returns a probability in the form,

$$\hat{P}(y_j = 1 \mid x_j). \tag{2.4}$$

Like deterministic classification, $\hat{P}$ is chosen from a family of probability functions P of a specific probabilistic classifier. Whilst, probabilistic classifiers map to a probability value in $[0, 1]$, they can still be seen as a mapping $X \to Y$ where Y is our event status since we can map the probability value onto Y via thresholding.

The deterministic and probabilistic settings are classification tasks. They can range from classical statistic methods such as logistic regression (Cox, 1958) to more modern machine learning methods such as kernel methods, ensemble methods and deep neural networks. Some classification methods, such as logistic regression, are naturally probabilistic, they can be converted to deterministic classifiers via thresholding. Others, such as support vector machines, are naturally deterministic classifiers. They can be converted to probabilistic classifiers via a method known as Platt Scaling Platt (1999).

2.2.1.2 Classification in a Time to Event Setting

Whilst classification as described above only takes the event status into consideration, we can also use classification for time to event modelling. The idea is rather than simply predicting the event, instead we aim to predict not only whether the event will happen but also whether it will happen within a defined time period. However, one issue that arises is censoring. For example, a unit may drop off the study before the event has happened and the observation time for the unit is shorter than the predetermined time period. Therefore, we must remove these units. We only keep units for which either the event happened within the defined period or where the event didn't happen in the defined period but the unit has been part of the study for at least the defined period. For example, if our aim is to predict whether a patient will relapse within five-years based on cancer patient annual check-up records after initial treatment we would subset our data in the following way. We keep records of patients who relapse within predetermined time - five years or those who have five-year records without relapse and we remove records of patients who stopped coming for check-up within five years or have less than five-year records. Once we take a subset of the data, we are able to consider this as a standard classification task.

2.2.1.3 Survival Models

In this section we will first begin by introducing survival modelling. We will then proceed to describe the two main functions central to survival models - hazard and survival functions. Survival models aim to predict whether an event will happen at a certain time for an individual unit given its covariates. Survival models usually model two types of functions, either the hazard functions or the survival functions. This is a framework introduced by (Cox, 1972). The definition of a hazard function is the probability that an event will happen at time t given that it has not yet happened before t. On the other hand, the aim of a survival function is to find the probability that an event has not yet happened before t, i.e. the unit has 'survived' until t.

Consider a unit i with time to event T_i and starting time $t = 0$, at time t we can model the probability that the unit will survive to time t by,

$$S(t|x_i) = P(T_i > t|x_i). \tag{2.5}$$

We denote the probability that the event will happen before time t by,

$$F(t|x_i) = 1 - S(t|x_i) = P(T_i \leqslant t|x_i). \tag{2.6}$$

The probability density function of the event will happen at time t is,

$$f(t|x_i) = F'(t|x_i) = -S'(t|x_i). \tag{2.7}$$

Most survival models are modelling the hazard rate which is the probability that the event will happen in the next time interval given that it has not yet happened before t,

$$\begin{aligned}
h(t|x_i) &= \lim_{\Delta t \to 0} \frac{P((t \leqslant T_i < t + \Delta t \,|\, T_i \geqslant t) \,|\, x_i)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t \leqslant T_i < t + \Delta t \,|\, x_i)}{P(T_i \geqslant t \,|\, x_i)\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(T_i < t + \Delta t \,|\, x_i) - P(T_i \leqslant t \,|\, x_i)}{P(T_i \geqslant t \,|\, x_i)\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{F(t + \Delta t \,|\, x_i) - F(t \,|\, x_i)}{P(T_i \geqslant t \,|\, x_i)\Delta t} \\
&= \frac{f(t \,|\, x_i)}{S(t \,|\, x_i)} \\
&= -\frac{S'(t \,|\, x_i)}{S(t \,|\, x_i)}.
\end{aligned} \tag{2.8}$$

From the hazard function 2.8 we can derive $S(t \mid x_i)$ and $f(t \mid x_i)$. First, we introduce cumulative hazard function, denote Λ,

$$\Lambda(t \mid x_i) = \int_0^t h(u \mid x_i) du, \tag{2.9}$$

then solving the above differential equation 2.8 we have,

$$S(t \mid x_i) = \exp\big(-\Lambda(t \mid x_i)\big). \tag{2.10}$$

From the equation 2.8, we have the following,

$$f(t \mid x_i) = h(t \mid x_i) S(t \mid x_i). \tag{2.11}$$

Survival models range from simple non-parametric models such as Kaplan Meier models (Kaplan & Meier, 1958), proportional hazards models (Cox, 1972) (parametric and semiparametric), accelerated failure time models (Wei, 1992a) to black box/machine learning models, such as survival random forests (Ishwaran et al., 2008) and Cox boosting (Binder & Schumacher, 2008). Apart from Kaplan Meier Models, all these models are able to incorporate covariates. Models, such as the Kaplan Meier estimator, compute the survival function whilst proportional hazards models compute the hazard function. The process of fitting survival functions via maximum likelihood estimators as well as the various other survival models are detailed in chapter 3.

2.3 A Review of Risk Prediction Models

In this section, a review of various approaches to Risk Prediction modelling is provided. Much of the risk modelling literature comes in the form of domain specific research, where each domain models specific 'events' of interest. Events of interest can be unforeseen events in a natural or social setting, examples include the default of a loan or the onset of a disease. Risk prediction models can be divided into two classes of models, classification (deterministic or probabilistic) and survival. The choice of model class is dependent on the specific goals and application setting. If we are simply interested in predicting which of the units will experience the event, it is only necessary to use a deterministic classifier. Likewise, if we are interested in finding out the probability of the event happening for each unit, it is appropriate to use probabilistic classification methods. Finally, if we wish to model the time in which the event of interest will

happen, we need to use survival models.

In a non-temporal setting, the most common way to predict risk is by defining risk prediction as a binary classification task in both a deterministic or probabilistic setting, where the label 1 is given if the event happens and label 0 is given if the event doesn't happen. Classification methods have been applied to a variety of risk prediction problems in a range of application domains. The choice of method is often dependent on the goals of the risk prediction model in question. In medical applications the focus is often on building interpretable models for which the potential limitations of the method are well understood. In addition, domain knowledge may mean that preconceived parametric assumptions can be made. In such cases, this will result in a risk prediction application where the focus is on one specific method. Examples include the use of logistic regression to lung cancer prediction (Cassidy et al., 2008) and linear discriminant analysis to colon cancer (Barrier et al., 2006). As well as interpretable models, sometimes we may need to identify the most important risk factors associated with the event of interest, in such cases it may be necessary to utilize methods that perform variable selection. Examples include Kooperberg et al. (2010) who use elastic net and LASSO methods to build risk prediction models by selecting the relevant genetic risk factors associated with cancer. However, in many other application settings, there is lesser need for interpretability and the aim is simply to build risk prediction models that are as accurate as possible. This enables us to apply more 'black box' algorithms where the mathematical properties and interpretability of the models are less understood. Examples of the application domains include Credit Risk (Khandani et al., 2010), Customer Churn (Tsai & Lu, 2009), (Xia & Jin, 2008) and Cancer Prediction (Kourou et al., 2015). Each of these applications shares a common approach whereby a wide variety of classification methods are applied to the risk of event and their performances are evaluated.

One of the most common problems associated with risk prediction in a classification setting is data imbalance. Data imbalance occurs when the number of units in one class is significantly greater than the number of units in the other classes. This can cause the classifiers to be biased towards the majority class. Burez & Van den Poel (2009) introduce weighted random forests and alternative accuracy measures to deal with data imbalance. Xie et al. (2009) build on the idea of weighted random forests to develop improved balanced random forests in order to help overcome data imbalance. However, whilst both approaches appear to perform well on their chosen datasets, they risk overfitting and are not very interpretable.

Survival models in risk prediction have been predominantly used in medical applications where the event of interest is often a onset of disease. As with the classification setting, there is a strong focus on building interpretable models. The choice of model is often dependent on the parametric assumptions we want to make. The most popular type of survival model used is the proportional hazards model. Proportional hazards models assume that the variables have a mul-

tiplicative effect on the hazard. Proportional hazards models can be fully parametric whereby a distributional assumption of the hazard is made or semi parametric where no distributional assumptions are required. Common example applications include prostate cancer predictions (Halabi et al., 2003) and cardiovascular disease predictions (Pencina et al., 2009). On the other hand if we want to assume that the effect of the variables accelerate or decelerate through time, we can apply accelerated failure time models (Datta et al., 2007). In this thesis we focus on Proportional hazards models due to their ability to allow for models with no distributional assumptions. The majority of all such survival analysis studies focus on discovering relationships between covariates and the time to event rather than building predictive models that work well out of sample. In addition, they use a limited range of performance measures to evaluate their performance. As with classification methods, there are also a number of 'black box' methods which may provide greater performance but are less mathematically understood/interpretable. Examples of such risk prediction applications include Fantazzini & Figini (2009) who apply survival random forests to credit risk and Binder & Schumacher (2008) who apply Cox boosting to cancer prediction.

2.4 Model Evaluation

In this section we focus on the issues of model evaluation. We first revisit the motivation behind model evaluation and then describe one of the most common ways of splitting the training set and test set. Finally, we give an overview of the different loss functions used in my thesis.

As denoted in section 2.2.1 Building a Supervised Learning Task, for a given risk prediction task there are a huge number of risk prediction models to choose from. Whilst it is possible to make parametric assumptions and focus on one model, often this is not desirable and therefore we must choose our model from a selection of models. In addition, even if we use just one model it may still be necessary to test the validity of the fitted model. Having chosen our selection of method(s), model evaluation involves two parts - firstly splitting the data into the training set and test set, and secondly choosing the appropriate measures i.e. loss functions.

2.4.1 K Fold Cross-Validation

As introduced in section 2.2.1 Building a Supervised Learning Task, fitting and evaluating models require splitting the data set of size N into a training set of size m and a test set of size $N - m$. After we fit the model to the training set, we need to validate the model on the

test data to determine whether our model is under-fitting or over-fitting or generalized fitting. The problem with simply splitting the data set into two parts is that we cannot guarantee the training set and the testing set have the same distribution. This will introduce a bias to the fitted model and cause large variance on validation of the model over testing set. So how to split the data into training set and testing set will affect the effectiveness and reliability of the model validation. The method of K fold cross validation (Stone, 1974) solves this problem.

The K fold cross validation splits the whole data set into K groups known as folds. Each of the K folds takes turns to be the testing set, the rest of the folds are the training set. This ensures each fold of the data set be a part of training set $K-1$ times and be testing set 1 time. The model is fitted K times and the average of model outputs is the final output.

2.4.2 Performance Measures and Loss Functions

In machine learning tasks, there is a wide spectrum of performance measures. However, there is no perfect performance measure which is suitable for all tasks and data sets since each performance measure focuses on different aspects of machine learning tasks. As detailed in section 2.2.1 Building a Supervised Learning Task, machine learning models are qualitatively evaluated using loss functions. The loss functions produce a value based on the modelled result and the reality in terms of event outcome. The closer the modelled result with the true outcome, the smaller the value/loss, or the greater the reward. As with the choice of risk prediction model class, the choice of the loss function will depend on the aim of our risk prediction task. The loss functions in risk prediction are broadly divided into two types – discrimination and calibration. Discrimination measures measure a model's ability to discriminant between the two classes of outcomes - the event happens or the event doesn't happen. Calibration measures seek to measure how close the probability of the event compared to the actual outcome. Calibration measures can only be applied in a survival or probabilistic classification setting. However, discrimination measures can be applied to both deterministic classification models and probabilistic classification models (via thresholding). Let us first begin by mathematically defining a loss function.

Consider a set G that contains a set of possible mappings, let Y be the output space. We define a loss function L in the following form,

$$\mathsf{L} : \mathsf{G} \times \mathsf{Y} \to \mathbb{R}. \tag{2.12}$$

The loss function is not just a performance measure but also the function to be optimised in

the process of training our model to best fit the training data in order to achieve optimal goals.

2.4.2.1 Deterministic/Discrimination Loss Functions

In this subsection we give an overview of the deterministic loss functions used in this thesis.

Classification Accuracy In the deterministic setting, the simplest performance measure is the classification accuracy, which is defined as follows,

$$\mathit{Accuracy} = \frac{\mathsf{NCP}}{\mathsf{N}}, \tag{2.13}$$

where NCP is the number of units which have been correctly predicted, N is the size of data set.

The higher the classification accuracy, the better the classifier has performed. Therefore, the classification accuracy is a 'reward' function rather than a loss function. However, one problem with the accuracy is that it struggles to cope with an imbalanced classification problem in which one category greatly outnumbers another. For example, there was one rare disease in UK which only 5% of the population get. If we have a model that identifies 1 in 5 patients among 100 participants, the model would produce an accuracy of 96%. However, in reality this model would not be of any use since it fails to identify 4 out of 5 who have the disease. Accuracy in this case is a misleading performance measure since it would not allow us to see the granular detail of the model's performance-for example the number of people with the disease who have been correctly diagnosed, this is especially relevant when the data is imbalanced.

Due to the drawback of accuracy in dealing with imbalanced data sets, we will move our focus on detecting the instances we are interested in, i.e. the positive cases. In the example of rare disease prediction, the positive case is the patient with disease. The measures recall R and precision P come into play. These were first introduced in the Message Understanding Conference (Sasaki, 2007). The recall R is defined as,

$$\mathsf{R} = \frac{\mathsf{TP}}{(\mathsf{TP} + \mathsf{FN})}, \tag{2.14}$$

where the true positive TP is the number of units with the positive status where the model correctly predicts the positive status, the false negatives FN is the number of units with the positive status where the model incorrectly predicts the negative status.

The precision P is defined as,

$$\mathsf{P} = \frac{\mathsf{TP}}{(\mathsf{TP} + \mathsf{FP})}, \tag{2.15}$$

where the false positives FP is the number of units with the negative status where the model incorrectly predicts the positive status. Back to our rare disease problem and the machine learning model, where recall is 0.2 and precision is 1 as there is no false positives, which means no health people is misclassified with disease. The recall of 0.2 is a bad result for a disease detection model as it only identifies 20% patients. If we go to another extreme, we predict all patients have the disease. We will have precision 0.05 and recall 1 as there is no false negative. We can see the improved recall is at cost of decreased precision. If the further medical examination is not costing, we will prefer high recall to minimize false negative cases to prevent life losses. On the other hand, if the disease isn't fatal and the further treatment is only for life improvement, but harmful and expensive, the false positive cases should be avoided so that a higher precision is desirable.

In the context of risk prediction modelling the precision expresses the proportion of instances where the event is predicted to have happened, did in fact happen whilst the recall expresses the proportion of the instances where the event happened for which the model correctly predicts to have happened. Although the both recall and precision focus on positives, in practical classification tasks, we have to trade-off between choosing which measure we maximize to improve our model. In some cases, we need find an optimal combination of recall and precision, the F1 score come into favour.

F1 Score: The $\mathsf{F1}$ score is the harmonic mean of the precision and recall (Sørensen, 1948) and defined as follows,

$$\mathsf{F1} = \frac{2 \times \mathsf{R} \times \mathsf{P}}{(\mathsf{R} + \mathsf{P})}, \tag{2.16}$$

where $\mathsf{F1} \in [0, 1]$. The closer F1 score is to 1, the better the performance of the model. When the F1 score is 1, the model has the perfect precision and recall.

F1 score doesn't count for true negative cases so it would be less useful in imbalanced data set where positive cases greatly outnumber. It is useful in our data sets where, as detailed in section 4.4.2 Demographic Statistics and Associativity Tests, the positive cases overwhelmingly

outnumber the negative cases. Another issue with the F1 score is F1 score isn't differentiable so it cannot be used directly as a loss function for optimization.

2.4.2.2 Probabilistic/Calibration Loss Functions

Whilst deterministic performance measures can be applied to both deterministic and probabilistic classifiers (via threshold), to evaluate the performance of our probabilistic classification models we require the probabilistic/calibration loss functions. Probabilistic performance measures can only be applied to probabilistic methods. The two loss functions used in this thesis are the logarithmic loss (log loss) and the Brier Score.

Log Loss: Logarithmic loss, cross entropy loss or simply log loss is first defined by (Cover & Thomas, 1991) and is a measure of how close the probability of obtaining the true event status is to 1. The log loss can be seen as negative log likelihood. From the output of a classifier, only the probability of the true event status will be used and taken into logarithmic domain. Since a probability is always within 0 to 1, the range of the log loss is $\mathbb{R}_-$. The closer the probability of the true event status is to 1, i.e. the closer to 0 the logarithmic probability, the better the model/classifier has performed. Therefore, the negative value of the logarithm is defined as Log Loss.

Let the set P_t contain a set of possible probabilities p_t of obtaining the true event status, we can define the log loss in the following way,

$$L_{logloss}: P_t \times Y \rightarrow \mathbb{R}_- \tag{2.17}$$

For an individual unit i, p_{ti} is the predicted probability of obtaining true event status for unit i, we calculate its log loss in the following way,

$$(p_{ti}, y_i) \rightarrow -\log\big(p_{ti}(y_i)\big), \tag{2.18}$$

$p_t(y_i) \in [0,1]$, the closer to 1 the p_{ti}, the smaller of $-\log(p_{ti})$, the closer the prediction to truth. For a smaller value of probability for the true event status, its larger log loss act like a penalty, the log loss is known as an incorrect prediction penalty.

In a binary classification setting, for a N unit data set, the log loss is,

$$LL = -\frac{1}{N}\sum_{i=1}^{N}\log\big(p_{ti}(y_i)\big). \tag{2.19}$$

As the log loss has the effect of penalising incorrect predictions, it is an optimal function to be optimized in model fitting process.

Brier Score: The Brier score (Brier, 1950) can be seen as a probabilistic equivalent of the mean squared error. It comes in the form of a squared difference between the predicted probability of the event of interest and the actual event status. If the event status is 1, the squared difference will be $(p_1 - 1)^2$, otherwise the squared difference will be $(p_1 - 0)^2$ where p_1 is the probability of obtaining status 1.

In a binary classification setting, let the set P_1 contain a set of possible probabilities p_1, we have the following Brier score mapping function:

$$L_{BrierScore}: P_1 \times Y \rightarrow [0,1]. \tag{2.20}$$

For an individual unit i, let p_{i1} be the probability of obtaining event status 1, the Brier score takes the form,

$$(p_{1i}, y_i) \rightarrow (p_{1i} - y_i)^2. \tag{2.21}$$

For a N unit data set, the Brier score is,

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_{1i} - y_i)^2, \tag{2.22}$$

The closer the p_1 to actual class status, the smaller the Brier score and the better fitted the model. If the Brier score is 0, the model is perfect. We can decompose the Brier score into two parts of calibration and determination (Stephenson et al., 2018). If we partition the probability $p_1 \in [0.1]$ into m mutually exclusive bins labelled by the index $k = 1, 2..., m$, denote n_k the number of units that have the probability of p_{1kj} fallen in the kth bin, where $j = 1, 2..., n_k$. The

Brier score can be rewritten in the following way,

$$\begin{aligned} BS = \frac{1}{N}\sum_{k=1}^{m} n_k \sum_{j=1}^{n_k} (p_{1kj} - \bar{y}_k)^2 \quad &\leftarrow \quad \text{calibration part} \\ + \frac{1}{N}\sum_{k=1}^{m} n_k\big(\bar{y}_k(1-\bar{y}_k)\big) \quad &\leftarrow \quad \text{discrimination part,} \end{aligned} \tag{2.23}$$

where $\bar{y}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} p_{1kj}$.

Wilcoxon Test: The Wilcoxon test (Wilcoxon, 1945) is a nonparametric test to compare whether there is a difference in location between two (paired) samples. It was first extended to model comparison by Demsar (2006). The Wilcoxon test for model comparison compares two models by ranking their loss functions from the same data set. In this thesis, we will use the Wilcoxon test to rank the Brier scores of the same data set from different prediction models and test whether the ordering of the Brier score results for one model is different from that of another.

The advantage of using the Wilcoxon test is that it does not make any distributional assumptions.

2.4.2.3 The Net Benefit via an Exchange Rate

All above measures evaluate the performance of risk prediction models from a purely statistical point of view. When we use a prediction model to produce a risk probability for an adverse event without taking other factors into consideration, we use 0.5 as the probability threshold to classify the units as at risk or not at risk. In other words, we put the same weight on harms and benefit. However, in a real world setting, a decision is often made after other practical factors are taken into consideration, i.e. making a decision involves trading off between harms and benefits .This provides the motivation for the Net Benefit introduced by (Vickers, 2016) as a performance measure. If a model tells us there are 10% units at the risk of the adverse event, we need take an action on them. In some circumstances, we may want to avoid missing false negative units at some cost, say we will exchange 10 actions to capture one true positive unit. This exchange rate means 9 false positive units worth one true positive unit. The exchange rate is derived from a trade-off between the consequences of action and no action. In our example the exchange rate is $\frac{1}{9}$.

The exchange rate Re also formulated by (Vickers, 2016) will affect the probability threshold

p_{td} which is used to convert the output of the probability model to deterministic classes in the following way,

$$R_e = \frac{p_{td}}{1-p_{td}} \tag{2.24}$$

This new analytic performance measure net benefit has gained a wide use in last decade because it incorporates the exchange rate/probability threshold in the following way,

$$\begin{aligned} \text{Net Benefit} &= \frac{TP}{N} - \left[\frac{FP}{N} \times R_e\right] \\ &= \frac{TP}{N} - \left[\frac{FP}{N} \times (\frac{p_{td}}{1-p_{td}})\right]. \end{aligned} \tag{2.25}$$

As an analytical measure, the net benefit weighs benefits and harms with a specific exchange rate. In almost any clinical setting we want to achieve the best health outcomes for the patients whilst at the same time being constrained by limited resources and budgets. In addition, some treatments may have adverse effects. The net benefit allows us to capture this trade-off by both rewarding true positives and also taking into account the costs of giving patients treatments which they may not require. In the context of diagnosing a disease, if the disease is fatal, early diagnosis is crucial. In the first step, we may use patients' demographic variables and a few blood test markers to predict the risk. We identify some patients at risk and in the next step we need a CT scan to confirm whether or not these patients have the disease. We know that using the risk prediction model in the first step alone we cannot guarantee not to miss any patients who have the disease, however, giving the CT scan to every patient can be both expensive and harmful. This provides the motivation for the use of the net benefit in decision making.

Assume there are 1000 patients in our study, there are 300 patients with disease. The prediction model A predicts 250 patents in high risk group who have the disease and 350 patients with false positive, yielding $F1 = 0.56$. The prediction model B gives 255 true positive and 300 false positive, producing $F1 = 0.59$. In terms of traditional measures, we would say the model B is better. If a doctor believes it would be worth testing 10 patients to identify one patient with the disease, it means the doctor put 9 times more weight on finding the disease at an earlier stage than avoiding a further examination, which can be translated to statistic probabilistic threshold $p_{td} = 0.1$.

Now we will use 0.1 as the probabilistic threshold instead 0.5 to re-run the models. This time

the model A produces $TP = 292$ and $FP = 480$, leading to net benefit 0.241. The model B gives $TP = 294$ and $FP = 700$, leading to net benefit 0.216. So for this doctor who will examine 10 patients to identify one patient with the disease, model A is a better choice as it gives a better clinical value. Intuitively we can see that model A only misses two more patients but brings down false positive cases by 220, which saves the cost for 220 unnecessary further examinations or treatments. The net benefit gives the quantitative measure for this cost and benefit trade off.

By varying the exchange rate, we are taking the harms and benefits into consideration to make our decision.

2.4.2.4 Survival Loss Functions

In this thesis we will focus on using two survival loss functions: the Concordance index and the integrated Brier score.

Concordance Index: The Concordance index (C-index) (Harrell et al., 1996) is a discrimination measure. The C-index is defined as the proportion of the pairs of units for which one unit is predicted to have its event first, really had its event first. In other words, the Concordance index is the measure of how well the model can rank the units based on their survival times not the probability of their survival times.

For a data set of size N, removing all censored units, we have N_r units. Our model predicts unit i will survive to time t_i, its really survival time is T_i, the $C-index$ can be defined in the following way,

$$C-index=\frac{\sum_{i,j=1}^{N_r} 1(T_i > T_j)\cdot 1(t_i > t_j)}{\sum_{i,j=1}^{N_r} 1(T_i > T_j)}, \tag{2.26}$$

where

$$1(u > v) = \begin{cases} 1, & \text{if } u > v \\ 0, & \text{otherwise} \end{cases}.$$

$C-index = 1$ corresponds to a perfect model prediction, and $C-index = 0.5$ represents a random prediction.

Integrated Brier Score: The Brier score we discuss in section 2.4.2.2 is for binary classifications settings. In the survival prediction setting, we need a measure of how well the models fit with reality over all times. The Integrated Brier Score measures the performance of a fitted survival function $\hat{S}(t)$ for the whole-time interval.

In a data set of size N, for unit i, we have covariate x_i, T_i is the time when the event happens and censoring is absent, the model predicts the probability of unit i survival at time t is $\hat{S}(t \mid x_i)$, so the Brier score for N units at time t is,

$$BS(t)=\frac{1}{N}\sum_{i=1}^{N}\left(1\,(T_i>t)-\hat{S}\,(t|x_i)\right)^2. \tag{2.27}$$

In reality, a survival analysis data set always has some units censored, for these units we only know the event has not happened after some time C(censoring time). We cannot simply use the above equation to calculate the Brier score for censored data. The standard way to overcome the issue of censoring is by using a weighting scheme independent of the survival model. This is known as Inverse Probability Censoring Weight (IPCW) introduced by Graf et al. (1999). The IPCW models the probability of censoring C, this enables us to put more weight on units that have not been censored. Let $G(t) = P(C > t)$ where G can be modelled using a survival estimator. Graf uses the Kaplan Maier Estimator, however alternative censoring estimators may be used (Gerds & Schumacher, 2008).

For unit i, let x_i be the covariate vector, $t_i = \min\{T_i, C_i\}$ and the event status δ_i. The formulation of the Brier score for the censored data is presented as follows,

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\left(0-\hat{S}\,(t|x_i)\right)^2\cdot 1(t_i \leqslant t \text{ and } \delta_i = 1)}{\hat{G}(t_i^-)} + \frac{\left(1-\hat{S}\,(t|x_i)\right)^2\cdot 1(t_i > t)}{\hat{G}(t)}\right), \tag{2.28}$$

where

$$1(u) = \begin{cases} 1, & \text{if u is true} \\ 0, & \text{otherwise} \end{cases}.$$

Gerds & Schumacher (2008) shows that such estimators are consistent.

To consider the performance of the model over the whole-time interval we can integrate over

the interval. Let t^* be the largest time to event or censoring in the data set,

$$\text{IBS}=\int_0^{t^*} \text{BS}(t)dt. \tag{2.29}$$

IBS is an overall measure for the prediction of the survival model. The smaller the IBS, the better the model fitted.

2.4.3 Regularization

There are multiple strategies in which risk prediction models can be improved. In this thesis our focus will be on one approach know as regularization. We therefore proceed to introduce regularization in the context of risk prediction models.

2.4.3.1 Bias and Variance trade-off

Before we go into the details of regularization, we introduce the concept of balancing bias and variance and the phenomenon of overfitting and underfitting.

As discussed in section 2.2.1 Building a Supervised Learning Task, we strive to find a map function to map input X to output Y. We assume there exists the relation between X and Y. We can write,

$$Y = g(X) + e. \tag{2.30}$$

g is the mapping function, e is error. In practices, we never make a perfect estimation of $g(X)$ and make e zero.

For a data unit i (x_i, y_i), we apply a modelled mapping $\hat{g}$ on x_i and get the estimated value of $\hat{g}(x_i)$, we have expected squared error,

$$\text{Error}(x_i) = \mathbb{E}\big[\big(y_i - \hat{g}(x_i)\big)^2\big]. \tag{2.31}$$

If we further decompose the Error, we can get the following,

$$\text{Error}(x_i) = \big(\mathbb{E}[\hat{g}(x_i)] - g(x_i)\big)^2 + \mathbb{E}\big[\big(\hat{g}(x_i) - \mathbb{E}[\hat{g}(x_i)]\big)^2\big] + \sigma_{ei}, \tag{2.32}$$

i.e.

$$\mathrm{Error}(x_i) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{IrreducibleError}. \tag{2.33}$$

The bias is the measurement how far the estimated value is away from the true value. The variance represents how far the estimated values spread out from the mean of the estimated values. Irreducible Error is the error that cannot be reduced by a good model and the reason for it can be the incompleteness or the variations of observations.

The data we acquire always contains the real information which represents the real pattern of data and some random noise, i.e. the variations of observations. We want the models only learn the real pattern of data. But in reality, when we train our models over a training set, we minimize the error and at the same time we may unintentionally force the models to pick up certain patterns of noise in the training set.

If we made the model too flexible to get the estimation too close to the true values in the training set, then this model has lower bias and high variance, leading to high error in the testing set. This is where the overfitting occurs. On other hand, if we use a simple model, for example, a linear model, to model non-linear data, we cannot capture the underlying pattern of data. This model may have a low variance, but has high bias, producing high error in both the training set and testing set. This model is underfitted.

In machine learning tasks, low bias and low variance are desirable but they simply don't go together. In general, decreasing the bias will increase the variance and decreasing the variance will increase the bias. There is a trade-off between bias and variance to avoid underfitting or overfitting. Finding a balance between bias and variance is always a challenge. Cross-validation can low bias and variance. Other methods, such as, bagging (Breiman, 1996), boosting (Freund & Schapire, 1997) and regularization (Tikhonov & Arsenin, 1977) can overcome overfitting. Regularization is a widely used technique in an attempt to solve the overfitting. In our future work we choose to focus on regularization for classification and survival models. We will therefore provide a brief introduction to regularization.

2.4.3.2 Regularization Methods

In machine learning settings, regularization is a process of modifying a learning model to make it more generalizable. Regularization actually introduces a penalty while the model is overfitting. This penalty is designed to penalize more overfitted features. The regularization used in our research primarily focuses on models where we can map the input X onto the output Y via coefficients β. They were originally developed for the regression setting by Tikhonov & Arsenin

(1977). It can be extended to a classification or survival setting via a regularized log likelihood LL_r. Let θ be the distribution parameters if applicable, a ridge regularized log likelihood can be defined in the following way,

$$LL_r(\theta, \beta) = LL(\theta, \beta) + \lambda \beta^T \beta. \tag{2.34}$$

λ is the tuning parameter which determines how much we want to penalize the flexibility of our model and especially penalize the overshot coefficients. If we chose $\lambda = 0$, we have no penalty at all and we just train our model by log likelihood. On the other hand, with the increase of λ, the impact of the penalty grows and coefficient estimates get smaller, even approaching to zero. So, finding an appropriate value of λ is a crucial step in regularization process. The mathematics of regularization can be seen via a trade-off of bias for a lower variance. In other words, adding the penalty of $\lambda \beta^T \beta$ to the optimization function of likelihood introduces a bias in the training process but will minimize the variance in the testing process so that overall error will reduce too. This ridge regularization penalty is also known as the L2 penalty.

While the ridge regularization constrains the absolute size of the coefficients to prevent overfitting and reduce the overall error of the models, sometimes we may want to select the most relevant variables, one way to do this is via LASSO regression (Tibshirani, 1996). LASSO is defined in the following way,

$$LL_l(\theta, \beta) = LL(\theta, \beta) + \lambda \|\beta\|_1 . \tag{2.35}$$

As well as imposing bias, LASSO also has the additional benefit of forcing some of the coefficients to equal zero. Forcing coefficients to be zero eliminates variables - therefore LASSO has the additional benefit of conducting variable selection. However, since an absolute value is not differentiable, LASSO likelihoods are computationally more difficult to optimize. Both Ridge and LASSO require the use of a parameter λ that needs to be tuned in the validation process. This in turn requires the choice of a loss function/performances measure. Further detail on the tuning process is provided in chapter 6 Regularization Problems in a Gym User/Patient Setting. The parameter λ reflects the size of the constraint being imposed. The LASSO regularization penalty is also known as L1 penalty.

Chapter 3

Weibull Full Likelihood Implementation

In this chapter, we detail a software implementation of a full likelihood estimation of survival models. The aim of this chapter is to implement the Weibull proportional hazards model to further develop our understanding of proportional hazards models - a class of models that are used extensively throughout this thesis. We first introduce the two main classes of survival models known as Proportional Hazards Models and Accelerated Failure Time Models. We then provide an overview with regards to how survival models can be fitted via maximum likelihood estimation and further derive the mathematical formulas of likelihood estimation. Finally, we proceed to specify the Weibull proportional hazards model and conduct a full implementation of the model based on the mathematical equation of its log likelihood. We test our model on a simulated data set and present the results.

The two main types of survival models are proportional hazards models and accelerated failure time models. Proportional hazards models assume the covariates have a multiplicative effect on the hazard function whilst accelerated failure time models assume the covariates have a multiplicative effect on the event time.

3.1 Proportional Hazards Models

The most commonly used survival models are known as proportional hazards models Cox (1972). Proportional hazards models aim to model a hazard function that we defined in section 2.2.1.3 Survival Models, as a function of covariates and time. A proportional hazards model consists of two components, at any time t, the baseline hazard function $h_0(t, \theta)$ and the

multiplicative function $G(\beta, x)$. The baseline hazard function $h_0(t, \theta)$ models the hazard rate independently of the covariates where $h_0(t, \theta)$ is a parametric distribution with parameters θ. The multiplicative function $G(\beta, x)$ incorporates the covariates via coefficients β,

$$h(t, \theta, \beta, x) = h_0(t, \theta)\, G(\beta, x). \tag{3.1}$$

A special case of proportional hazards models is the Cox proportional hazards model where $G(\beta, x) = \exp(\beta x)$,

$$h(t, \theta, \beta, x) = h_0(t, \theta) \exp(\beta x), \tag{3.2}$$

β is the vector of coefficients and x is the vector of the observed unit variables. In parametric models the $h_0(t, \theta)$ can be specified by choosing from a selection of distributions.

3.2 Accelerated Failure Time Models

Another alternative class of survival models to proportional hazards models are accelerated failure time models (Wei, 1992b). Instead of modelling the hazard, accelerated failure time models model the survival times T in the form,

$$\log T = -\beta x + \sigma W, \tag{3.3}$$

where W is an error term modelled by an appropriate distribution. In addition, $\exp(\sigma W)$ can also be denoted as the baseline time to event T_0. Whilst accelerated failure time models model the survival time, their survival and hazard functions can be derived to allow for fitting under the maximum likelihood estimation process as described in the next section. To derive the hazard and survival functions of accelerated failure time models we let T_0 be the baseline survival time in the form $\exp(\sigma W)$ and let $S_0(t)$ denote the baseline survival function and $T_0 = \exp(\sigma W)$, we can express $S_0(t)$ in the following form,

$$S_0(t) = P(T_0 > t) = P\left(W > \frac{\log t}{\sigma}\right). \tag{3.4}$$

Since we can denote $T = T_0 \exp -\beta x$, we can write the survival function in the form,

$$S(t, x) = P(T > t|x) = P\big(T_0 \exp(-\beta x) > t \mid x\big) = P\big(T_0 > t \exp(\beta x) \mid x\big) = S_0\big(t \exp(\beta x)\big). \tag{3.5}$$

We denote $S_0(t)$ as reference survival function. Using the relationship of equations 2.7 and 2.8 we can obtain the following relationships,

$$f(t, x) = -\big[S\big(t \exp(\beta x)\big)\big]' = f_0\big(t \exp(\beta x)\big) \exp(\beta x) \tag{3.6}$$

where $f_0(t) = S_0'(t)$,

$$h(t,x) = \frac{f(t \mid x_i)}{S(t \mid x_i)} = h_0\big(t\exp(\beta x)\big)\exp(\beta x), \tag{3.7}$$

where $h_0(t) = \frac{f_0(t)}{S_0(t)}$. We can see $S(t)$, $f(t)$ and $h(t)$ are all accelerated by the factor $\exp(\beta x)$.

3.3 Maximum Likelihood Estimation

One of the most popular ways of fitting survival models is by finding the appropriate β, θ that maximize the likelihood function as introduced by (Cox, 1972). For an individual unit i at time t_i, we denote the proportional hazard function to be,

$$h\left(t_{i,}, \theta, \beta, x_i\right) = h_0\left(t_{i,}, \psi\right)\ G(\beta, x_i). \tag{3.8}$$

3.3.1 Maximum Likelihood Estimation Derivation

For a given unit i, t_i is the time to event or censoring, whichever happens first, let ψ be the parameters of a parametric distribution θ and the vector of coefficients β if applicable, let $f(t_i, \psi)$ be the probability density function of the event happening at time t_i, we can write the likelihood function L in the form,

$$L_i(\psi) = f(t_i, \psi)^{\delta_i} S(t_i, \psi)^{1-\delta_i}, \tag{3.9}$$

where $\delta_i = 1$ if the event happens at time t_i or 0 otherwise. Using the relationship for a hazard function 2.11 as defined in section 2.2.1.3 Survival Models, we can rewrite the likelihood function in the form,

$$L_i(\psi) = h(t_i, \psi)^{\delta_i} S(t_i, \psi). \tag{3.10}$$

Assuming that each unit experiences the event or censoring independently of each other, for N units the likelihood function can be written in the form,

$$L(\psi) = \prod_{i=1}^{N} h(t_i, \psi)^{\delta_i} S(t_i, \psi), \tag{3.11}$$

by taking logarithms we can get the following result,

$$LL(\psi) = \sum_{i=1}^{N} \left\{\delta_i \log\big(h(t_i, \psi)\big) + \log\big(S(t_i, \psi)\big)\right\}. \tag{3.12}$$

Since $S(t_i, \psi) = exp(-\Lambda(t_i, \psi))$ 2.10 defined in section 2.2.1.3 Survival Models, we can rewrite equation 3.12,

$$LL(\psi) = \sum_{i=1}^{N} \{\delta_i log\big(h(t_i, \psi) - \Lambda(t_i, \psi)\big)\}. \tag{3.13}$$

The parameters ψ can be chosen to maximize log likelihood, where ψ includes the distribution parameters θ and the coefficients β if applicable.

The Maximum likelihood estimation method of fitting survival models can be used for both proportional hazards and accelerated failure time models. However, survival models can be fitted in alternative ways for example via boosting (Binder & Schumacher, 2008) and bagging (Ishwaran et al., 2008).

Computing the full likelihood requires the specification of underlying distribution for the baseline hazard for the proportional hazard models or the error term for the accelerated failure time models. If we do not want to specify a baseline we can use the semi parametric proportional hazards model, such as, the Cox proportional hazards. However, without a specified baseline model we are unable to use the full likelihood estimation as detailed above and instead we can use the partial likelihood. On the flip side, the use of the partial likelihood would only allow us to compute the hazard ratio, not the hazard itself. The partial likelihood method is described below.

The Partial likelihood: An alternative to fitting via the maximum likelihood is to fit the models via the partial likelihood (Cox, 1975). The advantage of this approach is that we do not have to specify a baseline survival distribution or hazard function.

Let us define a data set with size D in which all the units have independent time to event. At time t, no event has happened or censored, so all units are at risk. We call this data set as a risk set,

$$R(x, t_m) = \{m : t \geqslant t_m\}, \tag{3.14}$$

t_m is the time to event for the unit m and x is the covariate vectors of the risk set, its likelihood to failure at time t_m is,

$$\begin{aligned} L_m(\psi) &= P(\text{unit m fails} \,|\, \text{one fails from } R(x, t_m)) \\ &= \frac{P(\text{unit m fails} \,|\, t_m)}{\sum_{l \in R(x, t_m)} P(\text{unit l fails} \,|\, t_m)} \\ &= \frac{h(t_m \,|\, x_m)}{\sum_{l \in R(x, t_m)} h(t_m \,|\, x_l)}. \end{aligned} \tag{3.15}$$

In a proportional hazards model, $h(t,\theta,\beta,x) = h_0(t,\theta)exp(\beta x)$, this can be written in the form of ,

$$L_m(\theta,\beta) = \frac{h_0(t_m,\theta)\exp(\beta x_m)}{\sum_{l\in R(x,t_m)} h_0(t_m,\theta)exp(\beta x_l)}. \tag{3.16}$$

Cancelling top and bottom by $h_0(t_m,\theta)$ we get ,

$$L_m(\beta) = \frac{\exp(\beta x_m)}{\sum_{l=1}^{D}\exp(\beta x_l)}. \tag{3.17}$$

Assuming independence of each unit, the partial likelihood function for data set of size N therefore takes the form,

$$L(\beta) = \prod_{m=1}^{N}\left[\frac{\exp(\beta x_m)}{\sum_{l=1}^{D}\exp(\beta x_l)}\right]^{\delta_m}. \tag{3.18}$$

The partial likelihood is therefore the product of conditional probabilities of the event happening given the risk set at that time and that one event is about to occur. The advantage of partial likelihood estimation is that we do not have to make any distributional assumptions for the baseline hazard. However, since we only estimate the βs we are unable to compute the full hazard or survival functions which are required to calculate the loss functions such as the integrated Brier Score. On the other hand, the full likelihood models allow us to compute the survival function analytically once we have fitted the parameters and coefficients.

3.4 Weibull Proportional Hazards Model

In this subsection we will provide an overview of a specific example of a full likelihood model that we will implement.

The Weibull proportional hazards model (Lee & Go, 1997) is a fully parametric proportional hazards model that assumes a Weibull distribution for the probability density function. The Weibull proportional hazards model has been applied to a range of survival analysis applications ranging from engineering (Jardine et al., 1987) to agriculture (Caraviello et al., 2003).

The probability density function of the general Weibull distribution is denoted in the following form.

$$f_0(t,\alpha,\gamma) = \frac{\gamma}{\alpha}(\frac{t-\mu}{\alpha})^{(\gamma-1)}\exp\left(-(\frac{t-\mu}{\alpha})^{\gamma}\right), \tag{3.19}$$

where $\gamma \neq 1$ is the shape parameter, $\alpha \geqslant 0$ is the scale parameter and μ is the location parameter. Let $\mu = 0$, Set $\lambda = \alpha^{-\gamma}$, we can express the baseline probability density function of the Weibull proportional hazards model as,

$$f_0(t,\lambda,\gamma) = \lambda\gamma t^{(\gamma-1)}\exp(-\lambda t^{\gamma}), \tag{3.20}$$

From equation 2.7, we can integrate $f_0(t)$ to get $S_0(t)$,

$$S_0(t,\lambda,\gamma) = -\int_0^t f_0(u)du = \exp(-\lambda t^{\gamma}). \tag{3.21}$$

From equation 2.8, we can get the baseline hazards function of the normal Weibull distribution as follows, ,

$$h_0(t,\lambda,\gamma) = \frac{f_0(t)}{S_0(t)} = \lambda\gamma t^{(\gamma-1)}. \tag{3.22}$$

For unit i with observed covariate vector x_i, we have the following Weibull distribution hazard proportional function,

$$h(t,\lambda,\gamma,\beta \mid x_i) = \lambda\gamma t^{(\gamma-1)}\exp(\beta x_i), \tag{3.23}$$

Through the relationships of equations 2.9 and 2.10 we can derive the survival function,

$$S(t,\lambda,\gamma,\beta|x_i) = \exp\Big(\int_0^t h(u \mid x_i)du\Big) = \exp\big(-\lambda t^{\gamma}\exp(\beta x_i)\big). \tag{3.24}$$

and from equation 2.11, we can have the probability density function,

$$f(t,\lambda,\gamma,\beta|x_i) = h(t \mid x_i)S(t \mid x_i) = \lambda\gamma t^{(\gamma-1)}\exp(\beta x_i)\exp\big(-\lambda t^{\gamma}\exp(\beta x_i)\big). \tag{3.25}$$

From equation 3.12 and for a data set of N units we are then able to derive the associated log

likelihood in the following way,

$$\begin{aligned} LL(\lambda,\gamma,\beta) &= \sum_{i=1}^{N} \left\{ \delta_i \log\big(h(t_i,\psi)\big) + \log\big(S(t_i,\psi)\big) \right\} \\ &= \sum_{i=1}^{N} \left\{ (\delta_i) \left[\log\left(\lambda\gamma {t_i}^{\gamma-1}\right) + \beta x_i \right] - \lambda {t_i}^{\gamma} \exp(\beta x_i) \right\}. \end{aligned} \tag{3.26}$$

We then optimize this log likelihood function using the optim R function to find out λ, γ, β.

3.5 Simulated Data

Since we are implementing the full likelihood of a Weibull proportional hazards model, a data set is generated using a Weibull distribution with a predefined λ, γ and β. X_1, X_2 are vectors of size N and generated by normal distributions. Let x_i be a vector consisting of (X_{1i}, X_{2i}), Let β be a vector of (β_1, β_2), our aim is to create a data set of $\{t_i, \delta_i, x_i\}_{i=1}^{N}$ for N data units. The simulation formulation is based on the work of (Bender et al., 2005). We generate each time to event T_i using the following formula,

$$T_i = \left(\frac{-\log(U_i)}{\lambda \exp(\beta x_i)} \right)^{\frac{1}{\gamma}}, \tag{3.27}$$

while censoring time is computed in the following way,

$$C_i = \left(\frac{-\log(V_i)}{\lambda_c \exp(\beta x_i)} \right)^{\frac{1}{\gamma}}, \tag{3.28}$$

in which λ is modified,

$$\lambda_c = \lambda \left(\frac{1-r}{r}\right)^{\frac{1}{\gamma}}, \tag{3.29}$$

where both U_i and V_i are a draw from a uniform distribution of (0,1). The r is the event ratio, i.e. the ratio of number of units in which the event happened to the sample size.

The status $\delta_i = 1$ if $T_i \leqslant C_i$, 0 otherwise.

3.5.1 Simulation Experiment

In our experiment we set our λ to 0.2, γ to 0.8 and sample size N. We set the λ to be not equal to one to ensure that the data follows a Weibull distribution as opposed to just an exponential

distribution. The γ is set to less than one in order to ensure that the risk of the event on simulated data will decrease over time. We simulated the two variable vectors of X_1 andX_2, both are independently drawn from a normal distribution with size N. X_1 is simulated with mean=4, standard deviation 0.5. X_2 is simulated with mean 2 and standard deviation 0.5. We set β_1 to -5 and β_2 to 0.3, the event ratio r to 0.5, $x_i = [X_{1i}, X_{2i}]$.

We then applied our full likelihood implementation onto this data set to fit the λ, γ and β. We first set a data set of size N 10,000. Next, we simulate the data set 100 times, for each of the 100 data sets we fit the λ, γ and β by minimizing the negative log likelihood. Finally, we compare my mean fitted values, from the 100 data sets to the actual values that we set. We get the following result in table 3.1.

	λ	γ	β_1	β_2
True	0.2	0.8	-5	0.3
Mean	0.20558	0.79909	-5.00927	0.3036646
RMSE	0.024507	0.00987	0.0676	0.0276

Table 3.1: Results from a Simulated Data of Size $10,000$.

Table 3.1 demonstrates the mean fitted parameters are very close to the original parameter value we set. In order to demonstrate the impact of the sample size on the accuracy of the implementation, we extend this to a larger scale simulation for a given set of sizes (800, 1000, 2000, 3000, 5000 ,8000, 10000, 150000). We calculated the mean squared errors (MSE) between the parameters we set and the fitted parameters. We also calculated the variance of the MSE and plot both the MSE along with their error bars (Mean MSE $\pm$ standard deviation). The error bars provide a visual representation of the spread of the mean squared errors. We plot the log Sample Size (in order to get a higher resolution) to get the following plots.

Figure 3.1 shows the results and error bars for a fitted lambda. The error bar size and mean square error size decrease as the sample size increases.

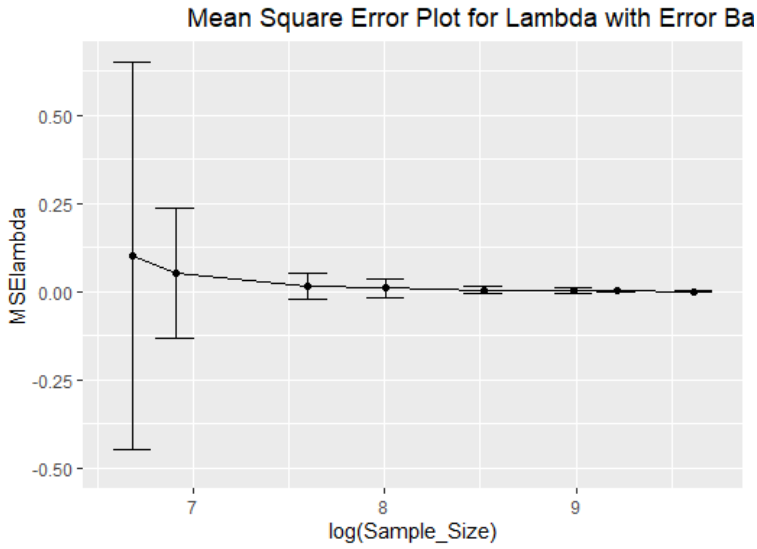

Figure 3.1: Plot for Fitted λ with Sample Size on the X Axis, MSE with Error Bars on the Y Axis.

Figure 3.2 shows the results and error bars for a fitted gamma. The error bar size and mean square error size decrease as the sample size increases.

Figure 3.2: Plot for Fitted γ with Sample Size on the X Axis, MSE with Error Bars on the Y Axis.

Figure 3.3 shows the results and error bars for the first fitted beta. The error bar size and mean square error size decrease as the sample size increases.

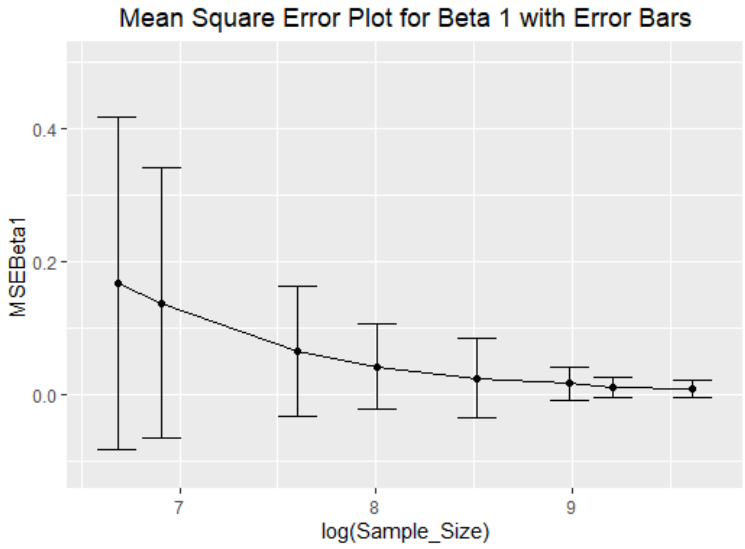

Figure 3.3: Plot for Fitted β_1 with Sample Size on the X Axis, MSE with Error Bars on the Y Axis.

Figure 3.4 shows the results and error bars for the second fitted beta. The error bar size and Mean Square Error size decrease as the sample size increases.

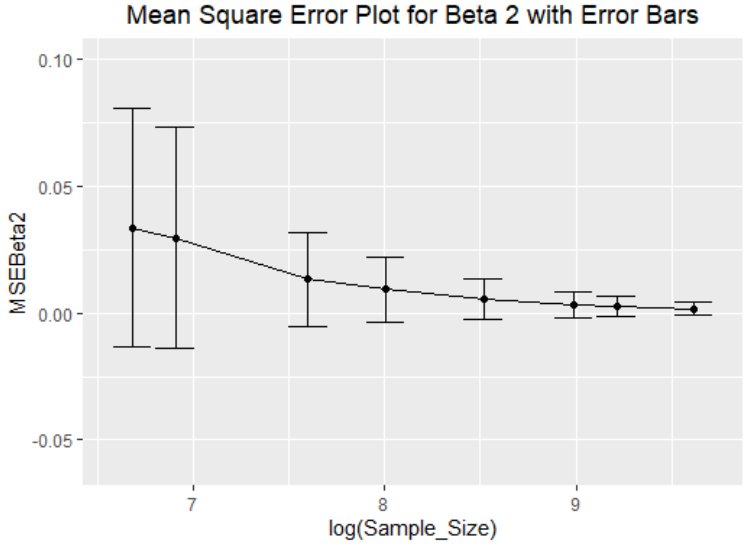

Figure 3.4: plot for Fitted β_2 with Sample Size on the X Axis, MSE with Error Bars on the Y Axis.

3.6 Summary

At first glance from the plots we notice that the mean MSE values are small. This indicates that the model fits the simulated data well as we expected and tells us that our model has been well implemented. Using our model, we can reveal the real pattern hidden in the data set. In addition, the mean MSE decreases with the increase of the sample size. This is consistent with the law of large numbers whereby the increased data size means that the fitted model gets closer to the expected value.

Another observation is that the size of the error bars decrease as the sample size increases, which indicates a decrease in variance. These phenomena can be seen in all plots and is consistent with the fact that the standard deviation is inversely proportional to the square root of the sample size. However, a notable feature is that when the sample size is small the lower bound of the error bars for many of the plots is below zero. This suggests that the standard deviation of the mean MSE is greater than the mean MSE itself. This implies that there is a huge amount of variation relative to the mean MSE itself. However, as we increase the sample size the lower bounds of the error bars are closer to zero. This indicates that variance decreases fast relative to the mean MSE as we increase the sample size.

Another area to take into consideration in terms of behaviour of the plots is the use of optimizers. This implementation uses the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) as its optimizer. This is the default optimizer in the 'optim' function. However, there are three other optimization algorithms available in the 'optima' function in R. Since the likelihood function is differentiable and therefore all optimizers can be used to fit its parameters. Different optimizers will have different convergence properties and will therefore impact how well the optimizing functions fits to the parameters.

The results show us that by maximizing the log likelihood optimized by a BFGS algorithm we are able accurately uncover the real pattern of the synthetic data with the Weibull distributions.

Chapter 7

Conclusion and Future Work

This chapter gives the conclusion to the thesis. We summarize our main findings from our investigations so far and give an outline for our future work

7.1 Summary of Findings

The thesis has the following main findings.

The first data set contains the gym user records which include the gym user demographic information and visit records. The data set inspire two sub investigations - *Predicting Gym User Churn* and *Predicting Gym User Visits*.

The second data set addresses patient records in rehabilitative care, where the patients are treated and helped back to work. The data set drives the two sub investigations - *Predicting Patient Treatment Times* and *Predicting Patient Treatment Types*.

Predicting Gym User Churn In our first investigation we look at whether machine learning modelling is able to predict whether or not a gym user will churn. We investigate firstly whether there is a link between demographic variables and a gym user churn. We find that based on exploratory analysis, demographic variables have a strong associativity with whether or not a user will churn. However, when we utilize the demographic variables

to build classification models to predict whether or not a user will churn, we find that we are unable to outperform a majority vote baseline. Even when we add further visit status information, we are still unable to outperform the baseline. We later conduct churn prediction in a survival setting and the models can outperform the baseline in terms of rank who will churn first. This sub optimal performances introduces the need for evaluating any predictive model against an majority vote baseline, to demonstrate its usefulness.

Predicting Gym User Visits Our second investigation looks at predicting individual visits in a moving window setting. When predicting individual user weekly visits, most machine learning models we build are slightly able to outperform a majority vote baseline over whether a user will visit in a given week with the user's demographic variables and previous visit records. We also find that there is a reduction in the predictability as we move forward in time since the available data set size decreases.

Predicting Treatment Times for Patient Rehabilitative Care The third investigation looks at whether we can use machine learning to design efficient treatment plans by predicting patient treatment times. This is the first study of its kind looking at modelling patient rehabilitative care as opposed to simply treat patients for a condition. Using associativity studies we first find that there is a link between certain demographic variables and treatment times. We then utilize risk prediction models that incorporate feature selection models to find the most relevant demographic variables. The feature selection models are able to reduce the number of variables that describe the model. Our treatment time prediction models are able to outperform the baseline from both a survival and regression settings. The results from the feature selection and the fact that demographic variables were able to outperform the majority vote baseline demonstrates that there is a link between demographic variables and patient treatment times. However, the weak outperformance of the risk prediction models demonstrates that there could be other factors that may determine how much treatment is needed. Additionally, the model performance could potentially be improved by reducing the amount of missingness in the categorical variables or creating interaction terms between the categories if deemed appropriate.

Predicting Patient Treatment Types Our fourth investigation progresses to extend the investigation into predicting treatment types. We treat the prediction of a treatment

type as a classification problem, using the demographic variables as our predictors. We find that due to the high-class imbalance, outperforming the majority vote baseline is difficult. However, in some cases the Adaboost algorithm was able to identify patients that required 'rare' treatments. Additionally, the model performance could potentially be improved by reducing the number of missingness in the categorical variables or creating interaction terms between the categories if deemed appropriate.

Methodological Contribution

1. **Comparison of tuning strategy performance of the F1 and Brier score LASSO regularization** Our first methodological contribution is to compare the tuning of the logistic LASSO via the F1 score and Brier score and demonstrate their merits via both the gym user churn data set and the patient treatment data set.

2. **Net benefit LASSO regularization** Introduces the net benefit LASSO and demonstrated its merits by showing that it is able to obtain greater net benefit performance.

3. **Integrated Brier Score regularization of Cox proportional hazards models** Demonstrates the benefits of using the integrated Brier score to tune regularized Cox proportional hazards models.

4. **Net benefit optimization via parameter tuning in machine learning models** Using support vector machines as an example we demonstrate how we optimize for the net benefit via modern machine learning methods.

7.2 Future Work

There are two possible areas for which my research can be extended. Firstly, investigate further the statistical properties of my methodological contributions by considering their performance on synthetic data. Alternatively, there are significant future areas of research that are beyond the scope of a PhD thesis. In the following subsection I explain possible approaches to generate synthetic data sets for both a classification and survival setting. However, any such further investigations would require more considerations.

7.2.1 Data Generation

In order to test the robustness of the models we can generate data sets with certain properties to evaluate the performance in these models. The main focus on this thesis will be imbalance, dimensions and censoring. This is motivated by the fact that the utility of loss functions can be dependent on the nature of data sets of interest. On the other hand, in the survival setting, varying proportion of censoring can affect the performance of model turning, this is especially of interest since the integrated Brier score account for censoring whilst the partial likelihood deviance and Concordance index do not. When generating the data sets we need to take into consideration what sort of data would most likely result in the method failing, in the context of parameter tuning this is heavily dependent on the loss function itself. From the perspective of the F1 and Brier score, the most obvious starting point would be the focus on imbalance. The Brier score is known to struggle to account for infrequent events therefore the focus will be on heavily imbalanced data. We now proceed to describe how we will generate the data sets in a classification and a survival setting.

7.2.1.1 Classification Data Set Generation

We aim to use a data set of fixed size 5000 but of varying dimensions. We will keep the number of meaningful variables fixed by drawing from a uniform distribution with mean 1 whilst drawing the rest of variables from a uniform distribution with mean value of 0. Also the data sets will be generated with different degrees of class imbalance to observe how loss functions perform on varying data sets. .

7.2.1.2 Survival Data Set Generation

For the survival data set generation, we also use a fixed data set size and number of meaningful but different dimensions. The data sets will be generated in the same way as we described in the section 3.5. Since the time to event and censoring are drawn from a Weibull Distribution, we will be confident that we are able to recover the coefficients with LASSO Cox regression.

We will generate different proportions of censoring data sets and study the performance

of our models with concordance index and integrated Brier score as turning measures for penalty parameter λ.

7.2.2 Future Areas of Research

The wellbeing research in this thesis, investigating the application of analytics to gym user behaviour and patient treatment planning, gives motivation for future research beyond the scope of this thesis. The following ideas are potential major areas to investigate, centred on optimising inter-related variables, such as specific outcomes:

1. **Reinforcement Learning Tool to Improve Gym User Well-being** Investigate the use of a reinforcement learning approach where the 'actions' to optimise different outcomes, such as user fitness, gym usage etc. The tool would test whether actions in the form of certain nudges or incentives can help maximize the reward in the form of better gym outcomes.
2. **Reinforcement Learning Planning for Hospitals' Patient Treatment** Investigate the use of a reinforcement learning approach where the 'actions' are the 'real-time' clinical decisions made to decide whether or not patients require treatment and where they should be allocated whilst the reward is the patient outcomes and hospital capacity.
3. **A Performance Index for Gyms and Hospital Care** An equivalent of the FTSE 100, the performance index for gyms and hospital care enables gyms and treatment centres to evaluate themselves against a benchmark. This will enable them to be assessed if they are underperforming or outperforming compared to their peers.
4. **National Comparison Tool for Gyms and Hospital Care**

   A ranking system that allows for different gyms and hospitals groups to be ranked base on optimising a 'group' of criteria. This would be a points-based system where points are given for a range of outcomes. The topic of research would be how to allocate the points for each possible outcome and how much weight to put on each positive or negative event.

Each of these ideas is of potentially interesting avenues to explore. However, since a large amount of data collection required over a long period of time, this would not be viable over the time scale of this thesis.

References

Akaike, H. (1971). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory*, 267–281.

Anderson, M. (2018). *Unhealthy employees now cost british firms six working weeks a year in lost productivitiy.* Mercer.

Barrier, A., Boelle, P.-Y., Roser, F., Gregg, J., Tse, C., Lacaine, D. B., ... Dudoit, S. (2006). "Stage II Colon Cancer Prognosis Prediction by Tumor Gene Expression Profiling". *Journal of Clinical Oncology*, *24*(29), 4685-4691.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating Survival Times to Simulate Cox Proportional Hazards Models. *Statistics in Medicine*, *24*(11), 1713-1723.

Binder, H., & Schumacher, M. (2008). "Allowing for Mandatory Covariates in Boosting Estimation of Sparse High-dimensional Survival Models". *BMC Bioinformatics*, *9*(14).

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... Jones, Z. M. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research*, *17*(170), 1-5.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144.

Breiman, L. (1996). "Bagging predictors". *Machine Learning*, *24*, 123-140.

Brier, G. (1950). "Verification of Forecasts Expressed in Terms of Probability". *Monthly Weather Review*, *85*, 1-3.

Burez, J., & Van den Poel, D. (2009). "Handling Class Imbalance in Customer Churn Prediction". *Expert Systems with Applications*, *36*, 4626-4636.

Caraviello, D., Weigel, K., & Gianola, D. (2003). Analysis of the Relationship Between Type Traits, Inbreeding, and Functional Survival in Jersey Cattle Using a Weibull Proportional Hazards Model. *Journal of Dairy Science*, *86*(9), 2984-2989.

Cassidy, A., Myles, J. P., van Tongeren, M., Page, R. D., Liloglou, T., Duffy, S. W., & Field, J. K. (2008). "The LLP risk model: an individual risk prediction model for lung cancer". *The British Journal of Cancer*(98), 270–276.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory 1st Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Cox, D. R. (1958). "The Regression Analysis of Binary Sequences". *Journal of the Royal Statistical Society*, *Series B (Methodological) 20*, 215-42.

Cox, D. R. (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society, Series B*, *34*(2), 187-220.

Cox, D. R. (1975). "Partial Likelihood". *Biometrika*, *62*, 269-276.

Datta, S., Le-Rademacher, J., & Datta., S. (2007). "Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO". *Biometrics Journal of the International Biometric Society*, *63*, 259-271.

Demsar, J. (2006). "Statistical Comparisons of Classifiers over Multiple Data Sets". *Journal of Machine Learning Research*, *7*, 1-30.

Dinh-Le, C., Chuang, R., Chokshi, S., & Mann, D. (2019, Sep 11). Wearable Health Technology and Electronic Health Record Integration: Scoping Review and Future Directions. *JMIR Mhealth Uhealth*, *7*(9), e12861.

Efroymson. (1966). Stepwise Regression–a Backward and Forward Look. *Eastern Regional Meetings of the Institute of Mathematical Statistics*.

Fantazzini, D., & Figini, S. (2009). "Random Survival Forests Models for SME Credit Risk Measurement". *Methodology and Computing in Applied Probability*, *11*, 29–45.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics*, *7*, 179-188.

Freund, Y., & Schapire, R. E. (1997). "A Decision-theoretic Generalization of on-line Learning and an Application to Boosting". *Journal of Computer and System Sciences*, *55*, 119-139. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/17240660/`

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gerds, T., & Schumacher, M. (2008). "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-censored Event Times". *Biometric Journal*, *48*, 1029-1040.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). "Assessment and Comparison of Prognostic Classification Schemes for Survival Data.". *Statistics in Medicine*, *18*, 17-18.

Halabi, S., Small, E., Kantoff, P., & et al. (2003). "Prognostic Model for Predicting Survival in Men with Hormone-refractory Metastatic Prostate Cancer". *Journal of Clinical Onocology*, *21*, 1232-1237.

Harrell, F. J., Lee, K., & Mark, D. (1996, 12). Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, 361–387.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hofmann, R. J. (1978). Complexity and simplicity as objective indices descriptive of factor solutions. *Multivariate behavioral research*, *13 2*, 247-50.

Horn, J. (1965). A Rationale and Test for the Number of Factors in Factor Analysis. *Psychometrika*, *30*(2), 179-185.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). "Customer Churn Prediction in Telecommunications". *Expert Systems with Applications*, *39*, 1414-1425.

Ishwaran, H., Udaya, U. B. K., Blackstone, E. H., & Lauer, S. M. (2008). "Random Survival Forests". *Annals of Applied Statistics*, *2*(3), 841-860.

Jardine, A. K. S., Anderson, P. M., & Mann, D. S. (1987). Application of the Weibull Proportional Hazards Model to Aircraft and Marine Engine Failure Data. *Quality and reliability engineering international*, *3*(2), 77–82.

Jenkins, J. (2014). *Sickness Absence in the UK Labour Market*. Office for National Statistics.

Kalra, D., & Ingram, D. (2006). Electronic Health Records. In *Information Technology Solutions for Healthcare* (pp. 135–181). Springer London.

Kaplan, E. L., & Meier, P. (1958). "Nonparametric Estimation from Incomplete Observations". *Journal of the American Statistical Association*, *53*(282), 457–481.

Khandani, A. E., Kim, A. J., & Lo, A. (2010). "Consumer Credit-risk Models via Machine-learning Algorithms". *Journal of Banking and Finance*, *34*, 2767-2787.

Kooperberg, C., LeBlanc, M., & Obenchain, V. (2010). "Risk Prediction Using Genome-wide Association Studies". *Genetic Epidemiology*, *34*, 643-652.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). "Machine Learning Applications in Cancer Prognosis and Prediction". *Computational and Structural Biotechnology Journal*, *13*, 8-17.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583-621.

Kubota, Y., Evenson, K. R., MacLehose, R. F., Roetker, N. S., Joshu, C. E., & Folsom, A. R. (2017). "Physical Activity and Lifetime Risk of Cardiovascular Disease and Cancer". , *49*, 1-23.

Lee, E. T., & Go, O. (1997). Survival Analysis in Public Health Research. *Annual review of public health*, *18*, 105-34.

Milošević, M., Živić, N., & Andjelković, I. (2017). "Early Churn Prediction with Personalized Targeting in Mobile Social Games". *Expert Systems with Applications*, *83*, 326-332.

Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Édouard Duchesnay (2018). *Scikit-learn: Machine learning in python.*

Pencina, M., D'Agostino, D. R. S., Larson, M., Massaro, J., & Vasan, R. (2009). "Predicting the 30-year Risk of Cardiovascular Disease: the Framingham Heart Study". *Circulation*, *119*, 3078-3084.

Platt, J. (1999, March). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". *Advances in Large Margin Classifiers*, *10*(3), 61-74.

Sasaki, Y. (2007, 10). The Truth of the F Measure. *PDF*, 1-10.

Semrl, J., & Matei, A. (2017). Churn Prediction Model for Effective Gym Customer Retention. *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*, 1-3.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*(5), 1–13.

Sørensen, T. (1948). "A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and its Application to Analyses of the Vegetation on Danish Commons". *Kongelige Danske Videnskabernes Selskab*, *5*, 1-34.

Stephenson, Coelho, C. A. S., & Jolliffe, I. T. (2018). "Two Extra Components in the Brier Score Decomposition". *Weather and Forecasting*, 752-757.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*, 111-147.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society*, *58*(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). "Sparsity and smoothness via the fused lasso". *Royal Statistical Society*, *67*, 91-108.

Tikhonov, A. N., & Arsenin, V. Y. (1977). "Solutions of Ill-Posed Problems". *Society for Industrial and Applied Mathematics*, *21*(2), 266–267.

Tsai, C. F., & Lu, Y. H. (2009). "Customer Churn Prediction by Hybrid Neural Network". *Expert Systems and Applications*, *36*(10), 12547-12553.

Tschuggnall, M., Grote, V., Pirchl, M., Holzner, B., Rumpold, G., & Fischer, M. J. (2021). Machine Learning Approaches to Predict Rehabilitation Success Based on Clinical and Patient-reported Outcome Measures. *Informatics in Medicine Unlocked*, *24*, 100598.

Vickers, A. J. (2016). "Net Benefit Approaches to the Evaluation of Prediction Models, Molecular Markers, and Diagnostic Tests". *BMJ*, *352*, 1-30.

Wallace, B. C., & Dahabreh, I. J. (2012). Class Probability Estimates are Unreliable for Imbalanced Data (and How to Fix Them). In *2012 ieee 12th international conference on data mining* (p. 695-704). doi: 10.1109/ICDM.2012.115

Wei, L. J. (1992a). "The Accelerated Failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis". *Statistics in Medicine*, *11*(14-15), 1871–1879.

Wei, L. J. (1992b). The Accelerated failure Time Model: A Useful Alternative to the Cox Regression Model in Survival Analysis. *Statistics in Medicine*, *11*, 1871–1879.

Wilcoxon, F. (1945, 12). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 80-83.

Xia, G., & Jin, W. (2008). "Model of Customer Churn Prediction On Support Vector Machine". *Systems Engineering-Theory and Practice*, *28*(1), 71-77.

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). "Customer Churn Prediction Using Improved Balance Random Forests". *Expert Systems with Applications*, 5445-5449.

Yuan, M., & Lin, Y. (2005). "Model Selection and Estimation in Regression with Grouped Variables". *The Journal Royal Statistical Society*, *68*, 49-67.

Zadeh, A., Taylor, D., Bertsos, M., Tillman, T., Nosoudi, N., & Bruce, S. (2020, May 15). Predicting Sports Injuries with Wearable Technology and Data Analysis. *Information Systems Frontiers*, *3*, 283–299.

Zeng, Z., Yao, L., Roy, A., Li, X., Espino, S., Clare, S., ... Luo, Y. (2019, sep 15). Identifying Breast Cancer Distant Recurrences from Electronic Health Records Using Machine Learning. *Journal of Healthcare Informatics Research*, *3*, 283–299.

Zhu, M., Zhang, Z., Hirdes, J. P., & Stolee, P. (2007). Using Machine Learning Algorithms to Guide Rehabilitation Planning for Home Care Clients. *BMC Medical Informatics and Decision Making*, *7*, 7-41.

Appendix A

Full Result Tables

	rn	1
1	(Intercept)	0.00
2	GenderFemale	0.00
3	GenderMale	0.00
4	Marital.StatusDivorced	0.00
5	Marital.StatusMarried	0.00
6	Marital.StatusSeparated	0.00
7	Marital.StatusSingle	0.00
8	Marital.StatusWidowed	0.00
9	Marital.StatusWith Partner	0.00
10	Education.LevelA Levels or NVQ Level 2/3	0.00
11	Education.LevelDegree	-0.06
12	Education.LevelDoctorate	-0.34
13	Education.LevelGCE's, GCSE's or NVQ Level 1	0.09
14	Education.LevelMasters	-0.04
15	Education.LevelNo Qualification	0.08
16	Education.LevelOther Qualification	0.10
17	Age	0.00
18	EthnicityAfrican	0.04
19	EthnicityCaribbean	0.00
20	EthnicityChinese	0.00
21	EthnicityIndian	0.00
22	EthnicityMixed White and Black African	0.00
23	EthnicityNot Stated	-0.36

24	EthnicityOther	-0.14
25	EthnicityOther Asian Groups	0.00
26	EthnicityOther Black Backgrounds	0.00
27	EthnicityOther Mixed Background	0.00
28	EthnicityOther White Background	0.00
29	EthnicityWhite British	-0.05
30	EthnicityWhite Irish	0.00
31	Diagnosis.CategoryDementia	0.00
32	Diagnosis.CategoryEpilepsy	0.00
33	Diagnosis.CategoryFunctional	0.00
34	Diagnosis.CategoryInflammatory/Infectious	0.00
35	Diagnosis.CategoryMovement Disorder	0.08
36	Diagnosis.CategoryMS	0.00
37	Diagnosis.CategoryNeuromuscular	0.00
38	Diagnosis.CategoryOther	0.00
39	Diagnosis.CategoryOther Tumour	0.00
40	Diagnosis.CategorySpinal Cord Injury	0.00
41	Diagnosis.CategoryStroke	0.00
42	Diagnosis.CategoryTBI	0.00
43	Diagnosis.CategoryVestibular	0.00
44	Diagnosis.SpecificAlzheimers	0.00
45	Diagnosis.SpecificAtaxia	0.00
46	Diagnosis.SpecificCervical	0.00
47	Diagnosis.SpecificEncephalitis	0.00
48	Diagnosis.SpecificFronto-Temporal	0.00
49	Diagnosis.SpecificGlioma - Grade 1	-0.17
50	Diagnosis.SpecificGlioma - Grade 2	0.00
51	Diagnosis.SpecificGlioma - Grade 3	0.00
52	Diagnosis.SpecificGlioma - Grade 4	0.00
53	Diagnosis.SpecificHaemorrhage	0.00
54	Diagnosis.SpecificHydrocephalus	0.00
55	Diagnosis.SpecificInfarct	0.05
56	Diagnosis.SpecificInflammatory - Other	0.00
57	Diagnosis.SpecificLumbar	0.00
58	Diagnosis.SpecificMeningioma - Grade 1	0.03
59	Diagnosis.SpecificMeningitis	0.00
60	Diagnosis.SpecificMinor - PTA <1hour	0.00

61	Diagnosis.SpecificMND	0.00
62	Diagnosis.SpecificModerate - PTA <1day	0.00
63	Diagnosis.SpecificNeuropathy - CIDP	0.00
64	Diagnosis.SpecificNeuropathy - CMT	0.00
65	Diagnosis.SpecificNeuropathy - Other	-0.27
66	Diagnosis.SpecificOther	0.00
67	Diagnosis.SpecificOther Brain Tumour	0.00
68	Diagnosis.SpecificOther Movement Disorder	0.00
69	Diagnosis.SpecificOther Neuromuscular	0.00
70	Diagnosis.SpecificPrimary Progressive	0.00
71	Diagnosis.SpecificRelapsing Remitting	0.00
72	Diagnosis.SpecificSarcoidosis	0.00
73	Diagnosis.SpecificSecondary Progressive	0.00
74	Diagnosis.SpecificSevere - PTA <7days	0.00
75	Diagnosis.SpecificSubarachnoid Haemorrhage	0.00
76	Diagnosis.SpecificVery Severe - PTA 7days+	0.00
77	Pre.Injury.Work.HoursFull Time	0.00
78	Pre.Injury.Work.HoursN/A	0.00
79	Pre.Injury.Work.HoursOff-Sick	0.00
80	Pre.Injury.Work.HoursPart Time	0.00
81	Pre.Injury.Work.StatusEmployed	0.00
82	Pre.Injury.Work.StatusEmployed Off Sick	0.00
83	Pre.Injury.Work.StatusEmployed on Graded Return	0.00
84	Pre.Injury.Work.StatusSelf-Employed	0.00
85	Pre.Injury.Work.StatusStudying	-0.22
86	Pre.Injury.Work.StatusUnemployed	0.45
87	Initial.Work.HoursFull Time	0.14
88	Initial.Work.HoursN/A	0.00
89	Initial.Work.HoursOff-Sick	0.00
90	Initial.Work.HoursPart Time	0.00
91	Initial.Work.StatusEmployed	0.00
92	Initial.Work.StatusEmployed Off Sick	-0.21
93	Initial.Work.StatusEmployed on Graded Return	0.00
94	Initial.Work.StatusMedically Retired	0.00
95	Initial.Work.StatusSelf-Employed	0.00
96	Initial.Work.StatusStudying	0.00
97	Initial.Work.StatusStudying Off Sick	0.00

98	Initial.Work.StatusUnemployed	0.00
99	Initial.Work.StatusVolunteering	0.00
100	Initial.Occupation.TypeClerical and Intermediate	0.00
101	Initial.Occupation.TypeEducation	0.00
102	Initial.Occupation.TypeMiddle or Junior Managers	0.00
103	Initial.Occupation.TypeModern Professional	0.00
104	Initial.Occupation.TypeRoutine Manual and Service	0.00
105	Initial.Occupation.TypeSemi-Routine Manual and Service	0.00
106	Initial.Occupation.TypeSenior Managers or Administrators	-0.26
107	Initial.Occupation.TypeTechnical and Craft Occupations	-0.23
108	Initial.Occupation.TypeTraditional Professional	0.00

Table A.1: LASSO Cox Model results

	rn	1
1	(Intercept)	295.68
2	(Intercept)	0.00
3	GenderFemale	0.00
4	GenderMale	0.00
5	Marital.StatusDivorced	0.00
6	Marital.StatusMarried	0.00
7	Marital.StatusSeparated	0.00
8	Marital.StatusSingle	0.00
9	Marital.StatusWidowed	0.00
10	Marital.StatusWith Partner	0.00
11	Education.LevelA Levels or NVQ Level 2/3	0.00
12	Education.LevelDegree	0.00
13	Education.LevelDoctorate	0.00
14	Education.LevelGCE's, GCSE's or NVQ Level 1	0.00
15	Education.LevelMasters	0.00
16	Education.LevelNo Qualification	0.00
17	Education.LevelOther Qualification	0.00
18	Age	0.00
19	EthnicityAfrican	0.00
20	EthnicityCaribbean	0.00
21	EthnicityChinese	0.00
22	EthnicityIndian	0.00

23	EthnicityMixed White and Black African	0.00
24	EthnicityNot Stated	0.00
25	EthnicityOther	0.00
26	EthnicityOther Asian Groups	0.00
27	EthnicityOther Black Backgrounds	0.00
28	EthnicityOther Mixed Background	0.00
29	EthnicityOther White Background	0.00
30	EthnicityWhite British	0.00
31	EthnicityWhite Irish	0.00
32	Diagnosis.CategoryDementia	0.00
33	Diagnosis.CategoryEpilepsy	0.00
34	Diagnosis.CategoryFunctional	0.00
35	Diagnosis.CategoryInflammatory/Infectious	0.00
96	Initial.Work.StatusSelf-Employed	0.00
97	Initial.Work.StatusStudying	0.00
98	Initial.Work.StatusStudying Off Sick	0.00
99	Initial.Work.StatusUnemployed	0.00
100	Initial.Work.StatusVolunteering	0.00
101	Initial.Occupation.TypeClerical and Intermediate	0.00
102	Initial.Occupation.TypeEducation	0.00
103	Initial.Occupation.TypeMiddle or Junior Managers	0.00
104	Initial.Occupation.TypeModern Professional	0.00
105	Initial.Occupation.TypeRoutine Manual and Service	0.00
106	Initial.Occupation.TypeSemi-Routine Manual and Service	0.00
107	Initial.Occupation.TypeSenior Managers or Administrators	0.00
108	Initial.Occupation.TypeTechnical and Craft Occupations	0.00
109	Initial.Occupation.TypeTraditional Professional	0.00

Table A.2: LASSO Generalized Linear Model results

Scoring	BL	Brier score	Log Loss	F1.score	NB R_e = 9/10	NB R_e = 7/2
lambda		0.426500	0.381600	0.196562	0.196562	0.998100
Predictors		β	β	β	β	β
(Intercept)		0.3374	0.0000	0.0000	0.0000	0.7352
Gender		-0.0129	0.0000	0.0000	0.0000	-0.1488
Single		0.8684	0.8444	0.7231	0.7231	0.9508
Partner		0.4449	0.3809	0.0000	0.0000	0.7415
Educatio Degree		0.0000	0.0000	0.0000	0.0000	-0.0814

GCE's, GCSE's or NVQ Level 1		-0.1314	-0.0727	0.0000	0.0000	-0.3318
Masters		0.0436	0.0000	0.0000	0.0000	0.9063
Education Qualification		0.0000	0.0000	0.0000	0.0000	0.7702
Age		0.0258	0.0271	0.0385	0.0385	0.0181
Ethnicity Other		0.0000	0.0000	0.0000	0.0000	0.1496
EthnicityWhite British		-0.4314	-0.3885	-0.0257	-0.0257	-0.6239
Diagnosis Functional		0.0000	0.0000	0.0000	0.0000	0.4582
Other		0.0000	0.0000	0.0000	0.0000	0.5009
Stroke		0.0000	0.0000	0.0000	0.0000	-0.0156
TBI		0.0969	0.0712	0.0000	0.0000	0.1809
Ataxia		0.0000	0.0000	0.0000	0.0000	-0.8844
Glioma - Grade 4		-0.2542	-0.1447	0.0000	0.0000	-0.7077
Infarct		0.0000	0.0000	0.0000	0.0000	-0.1513
Secondary Progressive		0.0000	0.0000	0.0000	0.0000	-0.6197
Work.Hours N/A		0.4456	0.4021	0.0798	0.0798	0.6374
Work.Hours Off-Sick		0.4794	0.4402	0.1299	0.1299	0.6319
Work.Hours Part Time		0.0000	0.0000	0.0000	0.0000	-0.0361
Employed on Graded Return		0.0000	0.0000	0.0000	0.0000	0.0310
Self-Employed		0.0000	0.0000	0.0000	0.0000	0.2022
Unemployed		0.0000	0.0000	0.0000	0.0000	0.0564
Student		0.0000	0.0000	0.0000	0.0000	-0.7406
Semi-Routine Manual/Service		-0.1057	-0.0373	0.0000	0.0000	-0.4357
Senior Managers/Administrators		0.2369	0.1464	0.0000	0.0000	0.9320
Technical/Craft		0.0000	0.0000	0.0000	0.0000	0.5138
Brier Score	0.126697	0.104407	0.105413	0.110099	0.110099	0.096007
Log Loss	4.376055	0.354212	0.353110	0.378319	0.378319	0.322475
F1 score	0.932367	0.932367	0.932367	0.933187	NA	NA
NB $\mathtt{R_e} = 9/10$	0.759276	0.759276	0.759276	0.759276	0.769186	0.763348
NB $\mathtt{R_e} = 7/2$	0.429864	0.469457	0.442308	0.427602	0.427602	0.533937

Table A.3: Full Results of Logistic Regression LASSO Regularization Classification for Treatment of Joint Meeting with an Occupational Therapist and a Psychologist Using Different Scoring Functions.

Scoring	BL	Brier score	Log Loss	F1.score	NB $\mathtt{R_e} = 9/10$	NB $\mathtt{R_e} = 7/2$
lambda		1.092100	1.082100	1.058200	1.089900	1.061200
Predictors		β	β	β	β	β
(Intercept)		0.1964	0.1384	0.2119	0.1558	0.1191
Male		-0.2740	-0.2737	-0.2718	-0.2718	-0.2718
Separated		-0.9020	-0.8995	-0.8921	-0.9020	-0.8932
Single		1.4226	1.4217	1.4316	1.4321	1.4321
Partner		1.0883	1.0848	1.0841	1.0925	1.0851
Education Degree		-0.1430	-0.1427	-0.1401	-0.1413	-0.1399
Doctorate		0.5756	0.5764	0.5825	0.5778	0.5822

GCE's, GCSE's or NVQ Level 1		-0.2480	-0.2494	-0.2534	-0.2463	-0.2524
Masters		1.3542	1.3538	1.3559	1.3573	1.3563
No Qualification		0.0565	0.0535	0.0442	0.0554	0.0453
Other Qualification		1.2060	1.2027	1.1922	1.2028	1.1931
Age		0.0234	0.0234	0.0244	0.0241	0.0244
Caribbean		0.6558	0.6382	0.5977	0.6532	0.6031
EthnicityOther		1.1699	1.1580	1.1417	1.1767	1.1457
Other Asian Groups		0.3410	0.3317	0.3174	0.3444	0.3203
White British		-0.7039	-0.7085	-0.7138	-0.7022	-0.7127
Epilepsy		-1.1963	-1.1905	-1.1733	-1.1927	-1.1749
Functional		0.8153	0.8161	0.8242	0.8206	0.8243
Diagnosis Category:Other		0.7879	0.7859	0.7865	0.7895	0.7869
Stroke		-0.2729	-0.2735	-0.2771	-0.2761	-0.2777
TBI		0.1762	0.1750	0.1711	0.1734	0.1712
Ataxia		-3.2973	-3.2806	-3.2464	-3.2909	-3.2500
Glioma - Grade 4		-1.0208	-1.0241	-1.0313	-1.0199	-1.0299
Haemorrhage		-0.5679	-0.5628	-0.5477	-0.5656	-0.5487
Moderate - PTA <1 day		0.7143	0.7077	0.6933	0.7185	0.6964
Secondary Progressive		-1.5358	-1.5513	-1.5841	-1.5359	-1.5792
Severe - PTA <7 days		-0.2761	-0.2749	-0.2716	-0.2748	-0.2715
Subarachnoid Haemorrhage		0.0197	0.0168	0.0153	0.0217	0.0164
Work.Hours: N/A		0.7861	0.7876	0.7973	0.7903	0.7967
Work.Hours:Off-Sick		1.0631	1.0623	1.0615	1.0649	1.0620
Employed on Graded Return		0.2297	0.2287	0.2253	0.2287	0.2255
Self-Employed		0.5453	0.5452	0.5445	0.5468	0.5449
Volunteering		0.4682	0.4575	0.4419	0.4758	0.4457
6th Form College A Levels		-2.3470	-2.3236	-2.2578	-2.3351	-2.2645
Account manager		-2.5678	-2.5407	-2.4569	-2.5418	-2.4638
Actor		-2.8627	-2.8396	-2.7918	-2.8658	-2.7997
Architect		-1.2141	-1.1937	-1.1344	-1.2014	-1.1406
Assistant Manager Pret		-3.2059	-3.1878	-3.1289	-3.1906	-3.1343
Barrister		0.1955	0.1956	0.1893	0.1880	0.1885
Business Operator, BT		-0.8607	-0.8304	-0.7635	-0.8624	-0.7738
CAMHS Maudsley NHS Trust		-2.0712	-2.0466	-1.9718	-2.0520	-1.9786
Carer		-2.7521	-2.7182	-2.6511	-2.7590	-2.6623
Cashier for sainsbrys		-0.7354	-0.7075	-0.6344	-0.7237	-0.6425
Creative advertising		-2.5840	-2.5585	-2.4754	-2.5604	-2.4820
Creative imaging		-0.9047	-0.8671	-0.7747	-0.8977	-0.7866
Events Planner		-4.8790	-4.8650	-4.8290	-4.8778	-4.8340
Finance director		-3.2508	-3.2320	-3.1829	-3.2437	-3.1885
Fitness coach		-3.1763	-3.1410	-3.0376	-3.1565	-3.0483
Garment technologist		-2.0502	-2.0212	-1.9511	-2.0471	-1.9611
GP		-2.4704	-2.4521	-2.4083	-2.4730	-2.4148
IT consultant		-2.3643	-2.3384	-2.2712	-2.3571	-2.2795
IT programmer		-1.1074	-1.0778	-1.0053	-1.0996	-1.0144

IT sales		-2.9380	-2.9098	-2.8491	-2.9343	-2.8570
Lecturer		-2.0015	-1.9764	-1.9337	-2.0138	-1.9427
Logistics		-2.3509	-2.3284	-2.2614	-2.3372	-2.2682
Lorry Driver		-2.0928	-2.0757	-2.0359	-2.0905	-2.0416
Medical Engineer		0.1462	0.0250	0.0000	0.1130	0.0000
Mnagement consultant		-2.7568	-2.7326	-2.6869	-2.7672	-2.6957
MSU Manager		-1.1737	-1.1530	-1.1117	-1.1749	-1.1182
Nursery School Teacher		-0.7760	-0.7353	-0.6370	-0.7674	-0.6498
Occupational Therapist		-4.3021	-4.2788	-4.2272	-4.3002	-4.2344
Operations manager		-3.7558	-3.7353	-3.6642	-3.7361	-3.6702
Owner of station news agent		-2.1584	-2.1388	-2.0950	-2.1604	-2.1017
Pa for local council		-3.7127	-3.6962	-3.6505	-3.7078	-3.6560
Part time Judge/barrister		-2.3157	-2.2943	-2.2544	-2.3232	-2.2619
Phlebotomist		-2.8159	-2.7966	-2.7318	-2.7970	-2.7374
Prison officer		-2.6581	-2.6314	-2.5710	-2.6566	-2.5789
Psychology Professor		-3.4194	-3.3984	-3.3737	-3.4337	-3.3809
Self-employed builder		-2.2803	-2.2555	-2.1992	-2.2780	-2.2069
Sells fish at Billingsgate		-2.3651	-2.3386	-2.2626	-2.3507	-2.2708
Socail media		-1.4285	-1.4121	-1.3546	-1.4072	-1.3592
Solicitor		-0.9386	-0.9230	-0.8977	-0.9503	-0.9040
Student		-2.2617	-2.2534	-2.2244	-2.2528	-2.2268
Supermarket sales Assistant		-3.5027	-3.4781	-3.4287	-3.5042	-3.4359
Teacher		-1.4958	-1.4808	-1.4458	-1.4932	-1.4504
Trainee Accountant		-1.2234	-1.1965	-1.1117	-1.1995	-1.1190
Employed nurseery teacher		-2.7601	-2.7454	-2.7096	-2.7558	-2.7137
Vodafone Account Manager		-1.7205	-1.7001	-1.6311	-1.7024	-1.6366
Warehouse supervisor		-2.3621	-2.3359	-2.2742	-2.3613	-2.2828
Community out-reach (PC typing)		-1.3598	-1.3400	-1.3006	-1.3606	-1.3065
Modern Professional		0.0417	0.0351	0.0173	0.0382	0.0191
Semi-Routine Manual/Service		-0.1964	-0.2026	-0.2195	-0.1994	-0.2179
Senior Managers /Administrators		1.4932	1.4916	1.4878	1.4938	1.4885
Technical/Craft		0.8548	0.8502	0.8399	0.8545	0.8414
Occupation Type		0.2867	0.2812	0.2680	0.2860	0.2700
Brier Score	0.420420	0.051338	0.051900	0.053284	0.051495	0.053113
Log Loss	14.521143	0.220462	0.220209	0.226229	0.220901	0.225727
F1 score	0.733840	0.973890	0.972549	0.977212	NA	NA
NB $\mathsf{R}_\mathsf{e} = 9/10$	0.201201	0.552102	0.552102	0.552102	0.556312	0.552102
NB $\mathsf{R}_\mathsf{e} = 7/2$	-0.891892	0.433934	0.432432	0.424925	0.432432	0.426426

Table A.4: Full Results of Logistic Regression LASSO Regularization Classification for Treatment of Joint Meeting with an Occupational Therapist and a Psychologist Using Different Scoring Functions with Over Sampling.

Scoring	Integration Brier Score	Concordance Index

lambda	0.025000	0.003125
Predictors	β	β
Married		0.117425
Single		0.243217
Partner		0.210232
Education Degree	-0.046122	-0.333235
Doctorate		-0.991823
"GCSE's or NVQ Level 1"""	0.048810	0.288388
Masters		-0.644016
No Qualification		0.414973
Other Qualification		0.427951
Age	0.004363	0.007319
Caribbean		-0.048696
Indian		0.214541
Ethnicity Not Stated		-0.534069
EthnicityOther		-0.724628
White British		-0.013915
Diagnosis Functional		0.167587
Movement Disorder		0.406847
Category Other		0.089022
TBI		0.136977
Vestibular		-0.109897
Cervical		-0.052134
Encephalitis		-0.064028
Glioma - Grade 2		-0.345284
lioma - Grade 4		0.201982
Infarct		0.273620
Meningioma - Grade 1		0.581516
Meningitis		0.194764
Moderate - PTA <1day		0.110569
Neuropathy - CMT		-0.723916
Other Neuromuscular		0.117804
Work.Hours:Off-Sick	-0.040072	-0.303811
Work.HoursPart Time		-0.383855
Work.StatusEmployed Off Sick	-0.242464	-0.437289
Self-Employed		0.137861
Studying		-0.108843
Unemployed		-0.390084
Accounts Supervisor		-0.281066
Administrator		-0.122361
Antiques consultant		-0.619301
Bar man		-0.074762
Barrister		-0.241088
Barristers Clerk		-0.008959

BBC radio 3 producer		-1.061993
Business Operator BT		-0.389021
Cashier		-0.181576
Cashier for sainsbrys		-0.852485
ellular and molecular medicine		-1.325722
Civil Servant		-0.721820
Director		-0.600532
Finace analysis for red cross		-1.917809
History and RE secondary teacher		-0.409723
Investment Banker		-0.267114
Lawyer		-0.422220
Manager in Translation Service		-1.409934
Maths Teacher		-0.096629
NHS Administrator		-0.207771
NHS England		-0.143974
Night club security manager		-0.007675
Office worker for BT		-0.372258
Optometrist		-0.196595
Pa for local council		-0.054183
PA in childrens Charity		-0.270512
Paediatric Nurse		-1.173057
Police - desk based		-0.073419
Police Officer		-0.335871
Post office clerk		-0.271123
Project Manager		-0.470095
Psychiatric SpR		-0.798691
psychologist		-0.121077
RBS Bank Admin r		-0.897531
Registrar		-0.930864
Researcher		-0.883406
Retail		-0.753238
Sales Assistant		-0.812008
Science teacher		-1.012103
Security Guard		-0.983622
Self emloyed TV Producer		-0.072055
ISenior manager		-0.062647
Solicitor		0.057359
Spindoctor		-2.001778
Student		-0.290812
Tax officer		-1.160007
Teacher		-0.391208
TV Producer		-1.604616
Waiter		-0.168476
Education		0.211352
Middle or Junior Managers		0.394766

Modern Professional		0.117476
Senior Managers or Administrators	-0.019382	-0.463803
Technical and Craft Occupations	-0.043927	-0.799344
Traditional Professional		0.354332
Integrated Brier Score	0.081284	0.082440
Concordance Index	0.621792	0.606795
Brier score at 500 minutes	0.450554	0.452633
Brier score at 1000 minutes	0.35643	0.358186

Table A.5: Treatment Time Predictions using Cox LASSO Model with Different Scoring Functions