A Comparative Assessment of Time-to-Event Models for Distribution Cable Networks

by

Jagbir Singh Kullar

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Mechanical Engineering
University of Alberta

Abstract

With rising demands in industry for reliable electrical cable distribution networks comes an inherent need for utility providers to know well the condition of the assets in their network. Heightened expectations from regulators and consumers require methods of reliability assessment to improve the processes adhering to performance-based ratemaking models.

The present work is a comparative assessment of survival modelling techniques in a real-world application. This work investigates the incorporation of class balancing methods in survival analysis models, and how to assess model performance in the context of time-to-event failure probability. Numerous modelling strategies are assessed in the context of calibration and discrimination performance, leading to the choice of a model that most accurately describes individualized asset survival and hazard functions.

A method is developed for a multi-stage approach to risk scoring and failure prediction for underground, medium-voltage, power distribution cable. The method applies data class balancing techniques and survival analysis models to accurately describe the survival probability of individual cables in a distribution network. Class balancing is applied to the failure state of the cables in the dataset using both under-sampling and over-sampling methods. The balanced data forms the input data for various machine learning survival analysis models and the performance of the models are compared against models trained using the original, highly unbalanced, dataset. The resultant survival models are evaluated for their efficiency, discrimination power and overall

goodness-of-fit to determine the best suited model given the data. Additionally, examination of cable parameters leads to the determination of the importance of cable properties and operating conditions that most significantly influence the model and the failure likelihood of cables. The information is used to generate individualized hazard and survival functions for specific cable instances with their unique covariate conditions.

The work concludes with a discussion of the implementation of the comparative assessment in the utility providers' data analytics and reliability processes, suggestions for the supplementary capabilities for the annual cable testing program, and recommends further work and strategies to create a more all-encompassing strategy for cable reliability analytics.

To Richelle,
my best friend and biggest supporter

Acknowledgements

Without the help and guidance of many people, this work would have never been produced.

Without my industry sponsor, this project would never have been possible. While industry leaders, their desire to further understand aspects of their business vision enabled me to undertake this work. They gave me access to both their information and their personnel, for which I am extremely thankful.

I wish to thank my project team. Tessa Ryan, P. Eng and PhD Candidate, for guiding me through the theory and applications of underground cables, and her endless knowledge of the industry. She offered her experience and expertise regarding underground asset reliability to me, for which I am very grateful. Also, a thanks to Mojtaba Yeganejou, PhD Candidate, for his advanced knowledge about machine learning and artificial intelligence which helped me start this project.

I am indebted to both Dr. Scott Dick and Dr. Michael Lipsett for their unwavering support throughout the entirety of the project. Dr. Dick patiently guided me through my inexperience and my problems to help me develop the skill set I have today.

As my supervisor, Dr. Lipsett gave me the opportunity to be a part of his highly regarded project team, believing in me when I didn't believe in myself. He saw the untapped potential I possessed and pushed me to be a better scholar and a better person. Despite his overloaded schedule, Dr. Lipsett, without hesitation, was available when I needed assistance. His high expectations kept me focused and helped me undertake this project little by little, day by day, to what you see today.

I am eternally grateful to my family. My mother has always encouraged me to pursue a higher education in hopes of fulfilling my limitless dreams; my father has blessed me with his endless wisdom throughout this journey; and my brother has helped me navigate the twists and turns of life through his kindness and generosity.

Finally, to my fiancée and soon-to-be wife, Richelle, without her unconditional love and support, I would not have been able to achieve even this small effort.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An underground electrical distribution cable network is founded on the basis of providing safe and reliable power to consumers [10]. While governed by regulatory bodies, understanding the maintenance and reliability of distribution network has both service costs and an opportunity cost of lost production for utilities providers. It is therefore necessary to have strategies to minimize the costs of distribution network maintenance and to maximize distribution network reliability [126].

Reliability is concerned with the components of a distribution network *functioning* as intended and the *probability of success* (not failing) over a specified *duration* [108]. Much effort has been expended in developing procedures to repair a fault once it has occurred. Further effort has gone into developing inspection techniques [101, 121]. But reliability analytics is the key step to effective preventative maintenance and a greater understanding of the ongoing condition of the distribution network.

This work presents a strategy for enhancing the understanding of reliability concerns of medium-voltage distribution cables and provides a means of classifying high-risk cables in an estimated time-to-failure manner.

The rest of this introductory chapter gives the scope of the problem of concern, the need for the work that was undertaken, and applications of the results. The chapter concludes with an overview of the rest of the chapters in the present work.

1.1 Background Situation

Underground distribution cable is regarded as the most complex asset in a distribution network. The condition of the cables is largely non-observable, being buried directly in the soil, with numerous damage modes, and no one all-encompassing test to observe cable health and operating conditions [4].

In a five-year span, damaged and failed underground cable contributed to approximately 70% of the experienced customer outages in the utility provider's distribution network. The replacement of cable is among the most expensive asset in the network [16]. Locating and removing failed cable segments, installing new cable, as well as downtime experienced by consumers all contribute to the cost associated with unplanned outages.

1.1.1 Nature of the Problem

While there exists a theoretical useful life for the cables that adhere to the so-called *bathtub curve* [137], some cables remain in service much longer than their specified end of life while others fail in the normal useful life stage.

Understanding the factors contributing to premature cable failure could significantly reduce the number of unplanned outages and consumer outage hours experienced, particularly in environments and operating conditions where it is hazardous or costly to conduct in-situ inspections. Underground cables operating in such environments demands remote diagnostic tools capable of accurately estimating remaining useful life and factors contributing to increased failure likelihood, leading to a more effective reliability understanding.

Utility providers therefore need reliability analytics strategies to make their distribution networks more reliable.

1.2 Objective of the Present Work

The present work describes a method of comparative assessment of survival modelling of real-world data, accompanied with data class balancing techniques. The method includes not only a technique for generating probabilistic assumptions of the survival of underground distribution assets, but also a method of identifying the variables that most significantly influence the failure likelihood of these assets.

Since most survival analyses are done for the classification of soon-to-be failed events [110], the focus is typically a singular survival model with high accuracy and a plethora of previously cited use. While the results generated may be of significance, other survival models exist that may provide greater predictive power [96] and are a more efficient approach for determining risk scores and failure probability attributed to the input data.

Relevant previous work and reliability analyses are discussed, a comparative method is proposed, simulations and results are presented, and the merits and limitations of the work are discussed with recommendations for future work.

1.3 Overview of Contents

The following chapters present a literature review of previous work, methodology for the investigation, implementation of the method and the results, conclusions, and recommendations.

1.3.1 Literature Review

The review of literature describes previous research efforts in supporting fields that contribute to the present work.

1.3.2 Methodology

The methodology chapter presents the scope of the present research, based on the gap between the problem and established solutions, and develops a methodology for solving the problem.

This chapter offers a set of hypotheses to be validated in the present work and discusses the various combinations of models used for addressing the problem. The methodology focuses on the specific issues of data class imbalance and variable importance for underground, medium-voltage, distribution cable data.

1.3.3 Implementation and Results

The implementation chapter describes how the hypotheses were validated. Specific methods are discussed for multiple combinations of class balancing and survival models. Identification of the most appropriate method of survival analysis for underground cable data, first by simulation, and then in performance evaluation is explained.

The discussion was supplemented with limitations of modelling methods used, appropriateness of testing procedures, and contributing factors to the survivability of the cables.

1.3.4 Conclusions

The chapter of conclusions examines the merits and limitations of the comparative assessment methods, assesses how well the methods addressed the stated problem, and summarizes the original contributions of this work.

1.3.5 Recommendations

The final chapter recommends areas of further development of cable reliability diagnosis methods for real-world implementations and ends with the speculation about the future of reliability assessment for distribution asset data.

Chapter 2

Literature Review

This chapter reviews reliability and survival modelling literature pertaining to time-to-event data, identifies issues with current modelling systems, describes previous solutions for survival modelling, and highlights the limitations of existing solutions related to the problem of survival analysis of underground cables.

After a review of general survival analytics and the need for such an analysis, the discussion focuses on survival modelling. Performance metrics used for survival analysis and functions to determine underlying distributions are reviewed. Methods and models for evaluating time-to-event survival data are discussed, as well discussion surrounding the limitations surrounding conventional modelling approaches. Time-to-event data, its properties, key indicators, and limiting features, are considered, with emphasis on data class balancing techniques.

2.1 Survival Analytics

This section discusses the foundational principles of survival analysis and the applications in which survival analysis is beneficial as well as understanding and identifying key characteristics of survival data.

2.1.1 The Need for Survival Analysis

Conventional classification and regression models are machine learning algorithms used for prediction, however, differ slightly in their output. Classification models [39] are used to predict discrete outputs whereas regression models [43] are used to predict continuous values. Both classification and regression models provide an excellent framework for discovering patterns in

datasets that can lead to actionable intel. However, an intrinsic problem associated with these models arise when the data collected also contains information about instances in which the timing of an event is of interest to the research objective- the censoring status of the instances [52]. In most real-world cases, the data that requires analyzing contains more than one independent variable. This adds another level of complexity to modelling the data in an informative fashion; standard approaches to data analysis take a univariate (one variable) approach and therefore do not encompass all the independent variables simultaneously [64], resulting in less informed decision making. Hence, the need for specialized techniques that considers both time and multiple independent variables.

Survival analysis is highly researched, with applications in a variety of fields such as engineering [38, 140], healthcare [19, 99], and economics [51, 80] in which a critical objective of these applications is determining when a particular event will occur considering it has not yet occurred. Survival analysis involves the consideration of the times between a fixed starting point and a terminating event, which is either an event occurrence, or the end of a study. The use of survival analysis models improves traditional models by allowing the survival to be assessed with consideration of multiple variables and offers insight into the strength of factors relative to their importance in the model.

2.1.2 Survival Data

Survival data best describes data that measures the time to some event [3]. In industrial applications, the event is failure of an asset or a component of an asset. Cox and Oakes [25] outline three requirements for understanding the time to an event: the start time, $t = 0$, of the study; a scale, measuring the passage of time; and a clearly stated definition of what constitutes an event like, e.g., failure of an asset that requires complete replacement of an asset.

Survival data consists of both dependent and independent variables. In all survival analyses, the dependent variable is composed of two pieces of information- the *survival time* and the *event state* [56], eliminating the possibility of producing qualitative measurements of, e.g., quality of an asset.

Hougaard [62] describes the uniqueness of survival data as stemming from responses being times unlike other variables whose responses are instantaneous and independent of response time. Consequently, during the time of the study, some assets do not fail under observation. These data are referred to as censored data.

Collet [23] defines censored data as instances that are only partially known, i.e., failure of an asset does not occur under observation. Most common to industrial applications is *right censoring*. If an asset has not experienced failure during the time of the study and its survival time is known throughout the duration of the study, it is assumed that the failure will take place after the observation window and the instance is *right censored* [87]. Clark et al. [22] and Jenkins et al. [71] characterize censoring status as a dichotomous indicator variable. The censoring state is denoted using binary classification with the value zero indicating the absence of the event and conversely, a value of one indicating the presence of a defined event.

Hougaard's [62] notion that assets that have failed, i.e., their exact lifetime is observed, contribute to the density function, $f(t)$, whereas if the censored assets contribute to the probability that the useful life exceeds the end of the study. That is, the observed end time for each asset instance is given by the value, X, where T is the failure time or age and C is the censoring time.

$$X = \min(T, C) \tag{2.1}$$

If the asset end time, X, is an observed failure rather than a right censored observation, the censoring status, δ, is given by Equation (2.2).

$$\delta = I\{T < C\} \tag{2.2}$$

Independent predictor variables, or *covariates*, are critical in determining the likelihood of an event occurring. Kalbfleisch [75] categorizes covariates as either time-independent or time-dependent. While covariates can provide explanatory insight that may determine an asset's survival probability, Verweij et al. [134] and others [54, 77] deem that not all covariates are significant to the asset's survival or failure.

Kalbfleisch et al. [75] mathematically explain survival data as $\boldsymbol{X}$, a $n \times p$ matrix, where n is the number of observations and p is the number of covariates. For observations, $\boldsymbol{x}_i \in \mathbb{R}^p$, Witten and Tibshirani [138] describe survival data as a triplet, $(y_i, \delta_i, \boldsymbol{x}_i)$, consisting of observations, $\boldsymbol{x}_i \in \mathbb{R}^p$, and their associated survival time y_i, and censoring status δ_i.

2.2 Performance Metrics

In order to gauge a model's ability to accurately represent and forecast data-based predictions, the model, along with the results, must be validated for the sake of accuracy and predictive performance [12]. While many performance metrics are readily available for use in regression and classification problems, survival analysis concerns itself with those metrics in which the presence of censored data is taken into consideration.

Concordance Index

The concordance index (C-index) [58] is a widely used discrimination measure for time-to-event models. The C-index can quantify correlations between event times and risk predictions so as to discriminate between early events and later occurrences [85]. In survival analysis applications, it is often desirable to compare the discriminative ability of various models and the resulting predictions made based upon the models; the C-index assesses the probability that a model generates a higher probability of event occurrence for a true event than for a non-true event, similar to that of the area under the ROC curve [85]. As a goodness of fit measure, the C-index provides a performance metric that can be reported across many modelling situations, where the data contains censored events.

The C-index is developed in depth in Harrell et al. [58] and applied in several regression based studies [57] and for multivariate analyses [59]. In survival analysis applications, the C-index provides a performance metric capable of enabling the comparison of models to determine robustness and operability, however, the C-index must be critically examined in models; Uno et al. determine an upward bias if the amount of censoring in the test data is large [128]. The application of the C-index is widespread, with usage in medical applications [35, 45], reliability

assessments [54, 123], and survival predictions [92, 134], while evaluating a variety of survival models.

Brier Score

In order to gauge the variance between the predicted and actual survival probability of a subject, the Brier score [14] is used. The Brier score is, in essence, a loss function that measures the accuracy of probability predictions [79]; the lower the reported value of the Brier score indicates a better survival prediction. The Brier score, however, is dependent on a single time point [55] thus, has been extended as a metric to assess the overall model at all times- the integrated Brier Score (IBS) (Equation 2.3). The IBS is defined as,

$$IBS = \frac{1}{\max_i(x_i)} \int_0^{\max_i(x_i)} BS(t)dt \tag{2.3}$$

where, the Brier score, $BS(t)$, is defined through the number of instances, N, estimated probability, f_t, and actual event status, σ_t:

$$BS = \frac{1}{N} \sum_{t=1}^{N} (f_t - o_t)^2 \tag{2.4}$$

Unlike the C-index, which a measure of the discrimination performance of a model, Park et al. [102] describe IBS as a measure of a model's overall performance that incorporates not only discrimination but also the level of calibration of a model. Gel et al. [50] found that IBS reduces hedging which is generally regarded as a desirable trait for a performance metric. The functionality of the IBS is two-fold, firstly, as mentioned, it provides both a goodness-of-fit measure to evaluate the model performance, and second, it provides a basis for a prediction error curve and further, a survival function to be built upon for survival analysis.

Verweij et al. [134] conduct model validation for the analysis of low voltage paper insulated lead cable (PILC) using the IBS method and conclude that the Cox proportional hazards model (Cox PHM) outperforms other models. The validation is concurrent with the R^2 metric results, showing that Cox PHM provides the greatest predictive power in this study. The IBS is also used in a comparative study of survival models including Cox PHM and random

survival forests (RSF) by Farhadian et al. [45]. The IBS, in conjunction with the C-index and error rate, provides strong evidence supporting the notion that RSF, in this study, outperforms Cox PHM for survival analysis.

2.2.1 Underlying Distribution Analysis

Frequently, it is important to understand and test whether the data in question is formulated from a specific distribution. If the test attempts to determine the agreement between a sample distribution and a theoretical distribution, then it is regarded as a goodness-of-fit test. The goodness-of-fit test helps determine whether there is a relation between studied covariates and if sampled data represents a population distribution.

The Kolmogorov-Smirnov test (K-S test) [89] is a non-parametric statistical approach for comparing cumulative distribution functions (CDFs) to determine agreement or disagreement between empirical and theoretical distributions [86]. The K-S test maximizes the absolute differences between the CDFs. Extensive research has been conducted on, and with, the K-S test in a univariate and multivariate fashion. Schröer and Trenkler [117] study the distributions of K-S test statistics where unequal sample sizes are present. D'Agostino and Stephens [28] and Justel et al. [74] provide extensive insight into the use and formulation of the K-S test and its statistical power. This is further developed in Razali and Wah [109] where the power of multiple statistical distribution tests, including the K-S test, is compared to test for data normality.

2.3 Survival Analysis Models

Many survival models exist that are capable of handling survival analysis and prediction tasks. The models of interest in this study learn from survival datasets, D,

$$D = \{(y_i, \delta_i, \boldsymbol{x}_i) \mid \boldsymbol{x}_i \in \mathbb{R}^p\} \tag{2.4}$$

as discussed in Chapter 2.1.

This study considers survival models that produce entire survival probability, $\hat{S}(t|\boldsymbol{x_i})$, curves $\{t, \hat{S}(t|\boldsymbol{x_i})\}$, for *all points* $t \geq 0$, that can be specified by a grouped distribution as well as extended to survival distributions on an individualized basis. This is pertinent to reliability analytics in which individualized survival probabilities for all time points are the desired output. Such a model allows the computation of a key statistic- an individual asset's expected survival time.

Cox Proportional Hazards Model

The Cox proportional hazards model (Cox PHM) was developed by Cox [25] in 1972 for medical applications [29] and is still among the most widely used models in reliability and time-to-event analysis [7, 123]. Cox PHM is a semi-parametric approach to investigating the survival time of an asset with respect to one or many covariates. The semi-parametric nature of Cox PHM does not require the baseline hazard function, $h_0(t)$, to be specified,

$$\mathrm{h}(\mathrm{t}, \mathbf{x}) = \mathrm{h}_0(\mathrm{t}) \exp\left(\sum_{\mathrm{i}=1}^{\mathrm{n}_1} \beta \mathbf{x_i} + \sum_{\mathrm{j}=1}^{\mathrm{n}_2} \gamma_\mathrm{j} \mathbf{x_j}\right) \tag{2.5}$$

thereby allowing the model to function without any assumptions made for the shape of the baseline hazard function. While not required, the baseline hazard function can be estimated using the hazard ratios generated from the modelling procedure, where the hazard ratio is determined using the ratio of two expected hazards, a and b, using regression coefficients, β.

$$\exp\left(\beta_{1(a-b)}\right) = \frac{h_0(t) \exp\left(\beta_{1(a)}\right)}{h_0(t) \exp\left(\beta_{1(b)}\right)} \tag{2.6}$$

While Equation (2.5) considers both time-independent, $\boldsymbol{x_i}$, and time-dependent, $\boldsymbol{x_j}$, variable. This study considers only time-independent covariates; therefore Equation (2.5) is reduced to Equation (2.7).

$$h(t, \boldsymbol{x}) = h_0(t) \exp\left(\sum_{i=1}^{n_1} \beta \boldsymbol{x_i}\right) \tag{2.7}$$

The functionality of Cox PHM is two-fold; not only does the model estimate the probability of an event of interest occurring, but it is also useful in determining covariates of significance and their respective relative importance on the occurrence of the event. Cox PHM has shown its versatility and interpretability in a range of survival applications.

In medical applications, Cox PHM is used by Farhadian et al. [45] for a cohort of patients undergoing coronary stenting to determine covariates of importance that may lead to, or have caused, major adverse cardiac and cerebrovascular events. Farhadian et al. find that Cox PHM yields similar covariates of importance to that of more complex machine learning techniques such as random survival forests (RSF). The dataset used by Farhadian et al. consists of approximately 50% right-censored samples which can generate lesser predictive accuracy [144], however the C-index of Cox PHM is 0.63, which is comparable to RSF, with a C-index of 0.65, for the given dataset. Recall, the C-index represents the model's discrimination ability of a time-to-event model to quantify correlations between risk prediction and event times. The results conclude that even with many right-censored samples, Cox PHM still performs similarly to models that are not affected by skewness in the data.

Cox PHM is also used as a predictive tool used to analyze power consumption of energy-intensive buildings by Gonzalez-Dominguez et al. [54]. Ten variables pertaining to building quality and condition are included in the model, with results outlining covariates of significance and their respective increase and decrease on the asset's hazard. Healthcare buildings with more than 10,000 users in the area were 124% more likely to exceed reference energy consumption. With this study, Gonzalez-Dominguez et al. propose different optimization strategies to reduce energy consumption of healthcare buildings based on the findings of Cox PHM.

Cox PHM has been used for determining ideal asset replacement times for low-voltage paper insulated lead cable (PILC) by Verweij, van Houwelingen, and Prein [134]. In the study, Verweij et al. leverage the multivariate nature of Cox PHM to analyze 17 covariates based on the technical specifications of PILC cables to optimize a replacement schedule from an asset management perspective. The results of Verweij et al. indicate that the length of cable and previous outages experienced in a specific cable are most significant to the increased hazard of a

cable and the cable network. A C-index of 0.77 is achieved with the data used in the study, which is in accordance with the average predictive accuracy of using Cox PHM [26].

Tang et al. [123] use Cox PHM to analyze a network grid of low, medium, and high-voltage cables to determine factors of significance that increase the likelihood of cable failure. The covariates in the study include the cable manufacturer and cable length. Tang et al. conclude that longer cables possess a greater risk that contributes to a larger failure probability. Further, the study determines that different manufacturers and their respective manufacturing processes also contribute to varying degrees of failure likelihood; all of which is determined solely and accurately, by Cox PHM.

Gradient Boosted Models

Gradient boosting and gradient boosted models provide a powerful framework for handling a vast range of objectives, with Friedman et al. developing an extension to gradient boosting for statistical estimation [48] and regression objectives [47].

Gradient boosting leverages multiple base leaners that, when used alone, are only slightly better than randomly guessing the outcome of a problem [115]. Base learners are simply regression estimators that are sequentially refined in an additive fashion to minimize a loss function [145], where the loss function is a measure of the deviation between predicted values from true value states. The use of multiple base learners effectively enhances the overall model's performance.

Gradient boosted models are used by Zhang and Haghani [145] to improve travel time predictions along freeways in the United States. Zhang and Haghani explore autoregressive integrated moving average (ARIMA), random forest, and gradient boosted models and conclude that the gradient boosted model is not only less sensitive to varying time outlooks while maintaining strong prediction accuracy, but also outperforms both ARIMA and random forests. Zhang and Haghani also confirm that gradient boosted models, by minimizing the specified loss function, reduce model bias and reduce variance as previously shown by Elith et al. [41].

He et al. [61] explore high-dimensional cancer data using a component-wise gradient boosted model to determine the effect on false discovery and model stability. Comparing various gradient boosted models used by Buhlmann and Yu [17] for high-dimensional data, Ridgeway [111] for boosting a proportional hazards model, and Li and Luan [83] for boosting non-linear function forms, He et al. discover that multivariate boosting, with emphasis on model stability, provides extremely low false discoveries, that is false positive and false negative results when compared with univariate and conventional boosting models.

Gradient boosting via the optimization of performance metrics such as partial log-likelihood, Brier score, and C-index is also examined in a study of high-dimensional, B-Cell Lymphoma conducted by Nguyen [98]. Gradient boosted models based upon the optimization of minimizing Brier score outperform models built to optimize partial log-likelihood and C-index, with testing errors of 0.29 for Brier score and greater than 0.5 for both partial log-likelihood and C-index. The ability for optimization of performance metrics using gradient boosted models provides model stability and accurate feature selection for high-dimensional data where traditional statistical methods cannot.

Random Survival Forest

Random survival forests (RSF) were developed as an extension of Breiman's [13] random forests (RF) as a classification and regression model for right-censored survival settings. RSF is a non-parametric ensemble learning method formed by averaging tree base learners where, in a survival setting, the base learners are binary survival trees [68]. The non-parametric nature of RSF enables a data-driven approach to survival analysis, independent of model assumptions. Unlike univariate approaches that are limited by overfitting and lack of convergence, RSF is constructed to mitigate these issues, even in the case of high dimensional data [35]. Where highly correlated covariates in the data are concerned, models with restrictive assumptions, namely Cox PHM, are not capable of dealing with the data, whereas the RSF model is specifically suitable for the analysis of such a dataset [66].

Ishwaran et al. [66] developed RSF and study the prediction accuracy of RSF using medical datasets. A varying number of covariates, ranging from 10 to 100 covariates, are used

among the eleven datasets in the study. Cox PHM is used as a benchmark test to determine the efficacy of RSF for different types of data. Ishwaran et al. conclude that RSF has the lowest prediction error in all examples, and with faster computing time. As the number of variables, and effectively, the amount of noise in the model increases, Cox PHM performs progressively worse whereas RSF remains stable and not susceptible to noise. The performance of RSF, however, is shown to decrease when the number of censored instances increases. Further testing of RSF for primary biliary cirrhosis of the liver by Ishwaran and Kogalur [67] indicate that the number of trees, when increased, provides a significant reduction in prediction error, and plateaus after 200 trees. The prediction error ranges from 16.0% to 17.5%, depending on the splitting rule used.

Numerous medical studies have been conducted using RSF including Miao et al. [92], who explore one-year mortality risk predictions of patients with cardiac arrhythmias. For the 10,000 patients in the dataset, four models are created. Two RSF models, one containing forty covariates and one containing fourteen covariates, are compared against two Cox PHMs with an equal number of covariates. Miao et al. conclude that both the high-dimensional and simplified RSF risk models perform better than both the high-dimensional and simplified Cox PHM, with prediction accuracies of 0.81 and 0.79 for RSF models, respectively. RSF is also used to determine covariates of importance for the survival of gastric cancer patients by Adham et al. [1]. The prediction error of the models was in the range of 29% and 32%, with age, tumor size, and metastatic status among the most important covariates that effect the prediction accuracy.

Applications of RSF have since extended outside of the medical domain, for use in reliability analytics and risk management by Frisk et al. [49] and Fantazzini and Figini [44]. Frisk et al. apply RSF to fleet management services for the lifetime prediction of vehicle batteries by analyzing 291 covariates from over 33,600 vehicles spanning across 5 markets. Throughout the analysis, RSF parameters such as the number of split variables, node size, and number of trees in the survival forest were optimized, with results indicting the optimal parameter selection yields an error rate of 15%. Fantazzini and Figini analyze small and medium enterprise credit risk ratings to compare RSF and logistic regression in terms of prediction accuracy. The area under the ROC-curve (AUC) is examined for both in-sample and out-of-sampling data, with results indicating that RSF outperforms logistic regression for in-sample

data, however, underperforms for out-of-sample data; AUC results for RSF in-sample and out-of-sample are 0.932 and 0.767, respectively.

Support Vector Machines

Support vector machines (SVMs) were developed by Cortes and Vapnik [24] as a learning method for analyzing high-dimensional data to produce a dichotomous outcome [63]. The development of SVMs was later extended by Vapnik [133] to support continuous regression outcomes and for use in survival analysis by Van Belle et al. [131] to maximize the C-index of the survival model. More generalized survival SVMs [46], formulated based on the foundations of Cortes and Vapnik [24], include the ranking approach [42, 131, 132], the regression approach [118], and a ranking and regression hybrid approach [130].

The ranking approach utilizes SVMs as a classification method to rank the risk of instances relative to each other, rather than determining survival time [130]. The objective of the regression approach, the survival SVM used in this study, is to find a function that estimates survival times as a continuous outcome and is built upon support vector regression [133]; Shivaswamy et al. [118] furthered the regression approach to include censorship in the SVM and SVR models [46].

Comparative studies of survival SVMs have been conducted by Pölsterl et al. [105] to determine prediction performance of survival in numerous medical datasets. The models compared in the study include kernel-based, linear, and simple survival SVMs compared against a baseline Cox PHM. Pölsterl et al. examine five datasets, with varying degrees of censoring for various medical datasets including AIDS, lung cancer, and coronary artery disease data. The C-index was used as one of three performance metrics, and the results indicated that survival SVMs performed on the same level of Cox PHM, with C-index values in the range of 0.68-0.76 among the various datasets.

Condition based monitoring and remaining useful life of machine components using SVMs is studied by Widodo and Yang [136]. The survival analysis encompasses SVMs to estimate the survival probability of failure time of bearing components in machinery to predict

failure times. The SVM used by Widodo and Yang accounts for censored data that is directly implemented into the prognostics modeling through the Kaplan-Meier estimator function. The predicted failure time compared with actual failure time of bearings was 98.51%, with a RMSE of 0.073, outlining the robust nature of SVMs when coupled with survival analysis.

Kernel based SVMs are used to study heart failure patients with correlation-based and ranking-based feature selection methods by Sujatha et al. [120]. Of the 299 patients in the dataset, AUC values between 0.85 and 0.88 and accuracy between 0.80 and 0.87 are achieved for the four kernel SVMs studied. Regression based kernel SVMs are compared against Cox PHM in Goli et al. [53] for a breast cancer study. Goli et al. report that linear kernel outperforms non-linear kernel regression SVMs using the Wilcoxon rank sum test [113] and performs greater than the Cox PHM. C-index values indicate that regression based SVMs perform better than Cox PHM, with values of 0.66 and 0.64 for the SVM and Cox PHM, respectively.

2.4 Handling Class Imbalance

Real-world datasets are often highly imbalanced; these imbalances can negatively skew the accuracy of predictions [33, 72, 77, 107, 127]. Imbalanced data is characterized by a difference in the number of instances, per class, of a variable. If the number of minority class instances and majority class instances significantly differ, the model becomes dominated by the majority class, with features of the minority class only slightly influencing the model [70]. Data imbalance can be dealt with using a data-level approach and algorithm approach [122]. The data-level approach involves re-balancing class imbalances in a preprocessing step and is the focus of resampling in this study.

Class imbalance, in the context of survival data, is commonly associated with the censoring status, with greater right-censored samples than uncensored samples. However, imbalances may also be prevalent in covariates, leading to bias selection of covariates in a model [18]. Data-level resampling techniques have been shown to improve classifier and regression model performance in a general and survival analysis application [36, 70]. Resampling techniques can be categorized into two sampling styles- under-sampling and over-sampling.

2.4.1 Under-Sampling Techniques

Under-sampling techniques reduce the cost at the learning stage of a model by removing instances in the majority class. Where cost-sensitivity is concerned, McCarthy et al. [90] and Liu et al. [84] show the benefit of under-sampling. This method, however, effectively discards potentially useful data that could be used in the model. Under-sampling may increase the variance of the classifier while also producing skewed probabilities [31].

Random Under-Sampling

Random under-sampling involves randomly removing random instances from the majority event class, i.e., the censored class, in the training data. This is often done until the number of samples of the majority class is equivalent to that of the minority class.

Random under-sampling can be utilized with various levels of sampling ratios, with a ratio equivalence showing a 30 percent increase in classifier performance when compared with a non-balanced dataset [147]. Yu et al. [143] determine a clear improvement in SVM accuracy when using random under-sampling for protein-ATP binding prediction. Dag et al. [30] examines one-, five-, and nine-year outlooks on predicting heart transplant outcomes on medical patients. Using, random under-sampling with various machine learning models, Dag et al. determines that under-sampling yields comparable results to more complex, over-sampling methods.

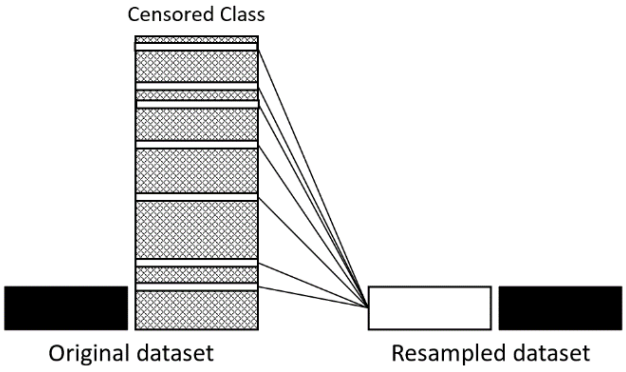


Figure 2.1: Random under-sampling data balancing

Tomek Links

Tomek Links (T-Links), an improvement on the nearest-neighbour rule [40], provide an alternative to random under sampling by removing pairs, from the majority class, of instances of opposites classes that are their own nearest-neighbour [125]. Removing the nearest-neighbour from the majority class in a dataset creates increases the space between the two classes in the training dataset, allowing for a more distinguishable and less noisy classification.

While T-Links do not provide a true class balance, results from a highly imbalanced dataset indicate that model accuracy is significantly improved when used with classifiers such as SVMs, Artificial Neural Networks, and Random Forests [40]. T-Links, as a noise removal tool, can also be coupled with other resampling techniques to improve classification accuracy [97]. A two-stage resampling method reduces the likelihood of information loss while still enhancing the robustness of a model.

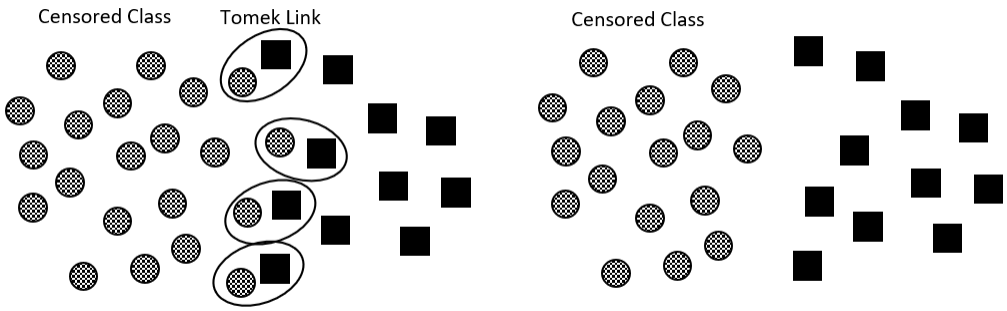


Figure 2.2: Tomek Links data balancing

NearMiss

NearMiss is an under-sampling technique that eliminates instances in the majority class based on their average distances from the minority class in the data space; this is done by removing, from the majority class, nearest-neighbours [60]. This technique preserves information from the training dataset, a common complication when using random under-sampling.

NearMiss is explored to great lengths by Mani and Zhang [88]. Three methods of NearMiss are used in which the number of opposite instances is eliminated to varying degrees.

The NearMiss methods in [88] are compared with random selection sampling in a kNN model. With an increase in the number of samples eliminated, the recall decreases and precision increases. The NearMiss method in which all opposite nearest-neighbour classes are eliminated from the dataset outperforms all other methods in the study. Classification models are compared by Oladunni et al. [100] in which highly imbalanced COVID-19 data is examined using NearMiss. Several NearMiss approaches are used similar to those in Mani and Zhang [88]. The results indicate that Boosted and Random Forest models present the greatest accuracy when under-sampling is used.

2.4.2 Over-Sampling Techniques

Over-sampling data involves introducing additional data to the minority class. When examining small datasets, over-sampling has proven to be an effective method to increasing the performance of a model [36]. Over-sampling outperforms under-sampling techniques where relatively low-dimensional data is present [6].

Random Over-Sampling

Random over-sampling is a non-heuristic approach to handling imbalanced data. In this technique, the minority class in the training data is duplicated randomly one or more times until both classes are equivalent.

In highly imbalanced datasets, Barandela et al. [5] showed that random over-sampling of the minority class increases the accuracy of predictions. While random over-sampling does not provide the highest accuracy in software defect prediction, the results from Bennin et al. [8] indicate improvement of failure-prone software module detection when compared against a non-balanced dataset. Random over-sampling, however, increase the likelihood of overfitting due to the duplication of samples in the minority class [78, 112, 135].

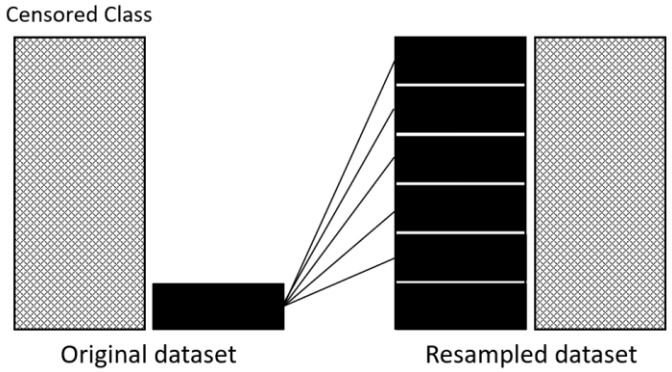


Figure 2.3: Random over-sampling data balancing

Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE interpolates synthetic instances along a line segment which connects randomly chosen data points in the minority class. This is done by computing k-nearest-neighbours for the creation of the sample [20, 107]. SMOTE overcomes the overfitting phenomenon created by random over-sampling and is focused on the creation of synthetic samples, rather than duplicate instances [20].

Chawla et al. [20] demonstrate improvement of accuracy for real-world data classification using SMOTE compared to other sampling methods. In a large-scale comparison, Pecorelli et al. [103] indicate that SMOTE outperforms several other over-sampling methods. SMOTE can also be used in a multi-class balancing fashion with a random forest classifier, with results indicating an improvement in the accuracy of the model [9]. Ishaq et al. [65] integrate SMOTE with various classifiers and found that classifier accuracy, precision, and recall results were strengthened in medical survival analysis.

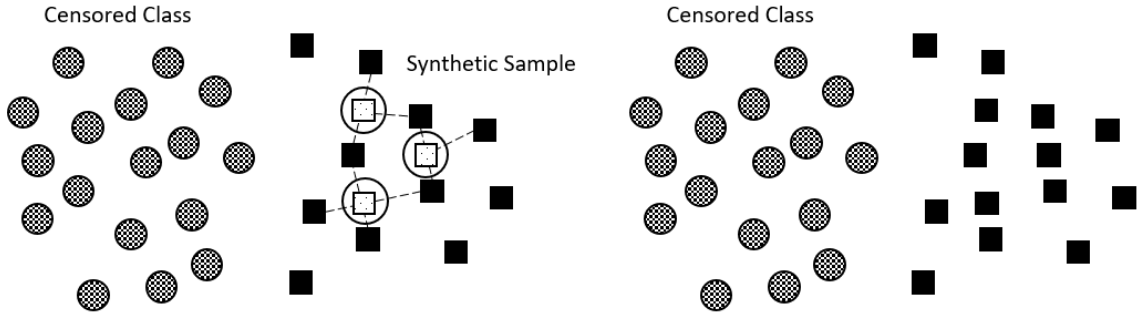


Figure 2.4: SMOTE data balancing

2.5 Summary of Previous Work

This chapter has given an overview of the theory and methodology pertinent to multivariate survival analysis. Biased performance metric results stemming from imbalanced data used in a survival analysis setting are of greatest concern, with specific attention given to data pre-processing and modelling methods that can achieve high accuracy while minimizing computational cost, false negative results, and skewed metrics.

Previous work conducted in the area of survival analysis, while advancing the expertise in the field, have been primarily conducted for medical applications, with minimal application in the research of underground power distribution cables. Jamet et al. [69] predict transplant eligible patients using RSF modelling methods, Bohannan et al. [11] use RSF to identify biomarkers in high-risk leukemia patients, and González-Domínguez et al. [54] use Cox PHM for predictive analysis of healthcare building energy consumption. All of these studies focus on a single survival model, used to generate information pertinent to prediction of medical related information.

Few researchers have applied survival analysis techniques to underground cable assets, however the datasets used in these studies are carefully curated such that large imbalances do not exist. Tang et al. [123] use Cox PHM to understand variables of significance that contribute to cable failure. The data used in the study is handpicked to create a dataset containing equal numbers of cables of various lengths. Verweij et al. [134] develop replacement strategies of PILC cable formulated through a Cox PHM survival analysis. Data used in this study is algorithmically selected based on location and topology. In the study, the IBS is used as the performance metric of choice.

Few studies encompass real-world, multivariate, and highly censored data. A highly censored dataset also limits researchers' ability to obtain true results, without inherent bias due to censorship.

Limitations of Underground Cable Survival Analysis

Current underground cable analyses have been limited to simple methods such as Cox PHM, in a multivariate fashion. In these studies, data balancing has been manually conducted, typically in the form of under-sampling whereby the data instances are selected from a master dataset in attempt to balance the number of censored and non-censored cases.

Further, the current underground cable studies are limited in the number of samples used in the study, and the number of covariates included. This is due, in part, to the limited amount of data obtainable, and the lack of covariate data that is included when initially forming the dataset.

To the best knowledge, no current survival analysis study has been conducted with the objective of a comparative analysis of various modelling methods with the inclusion of data class imbalance techniques for underground cable assets.

Chapter 3

Methodology

This chapter develops techniques for applying multivariate survival analysis techniques to medium-voltage underground cables with the inclusion of data class balancing techniques. The hypotheses to be validated are presented, a testing methodology is proposed, encompassing the process outlined in Figure 3.1, and an overview of the theoretical development required to fulfill that methodology is presented.

An examination of survival analysis modelling methods led to the selection of several models to be used in the study. Similarly, the selection of class balancing techniques that best balances data are applied concurrently with the survival models. The effects of class balancing were considered with respect to predictive accuracy and overall model performance. The limitations of commonly used models were uncovered, and a strategy was developed for using survival analysis techniques to create a robust technique, capable of developing individualized survival curves for underground cables.

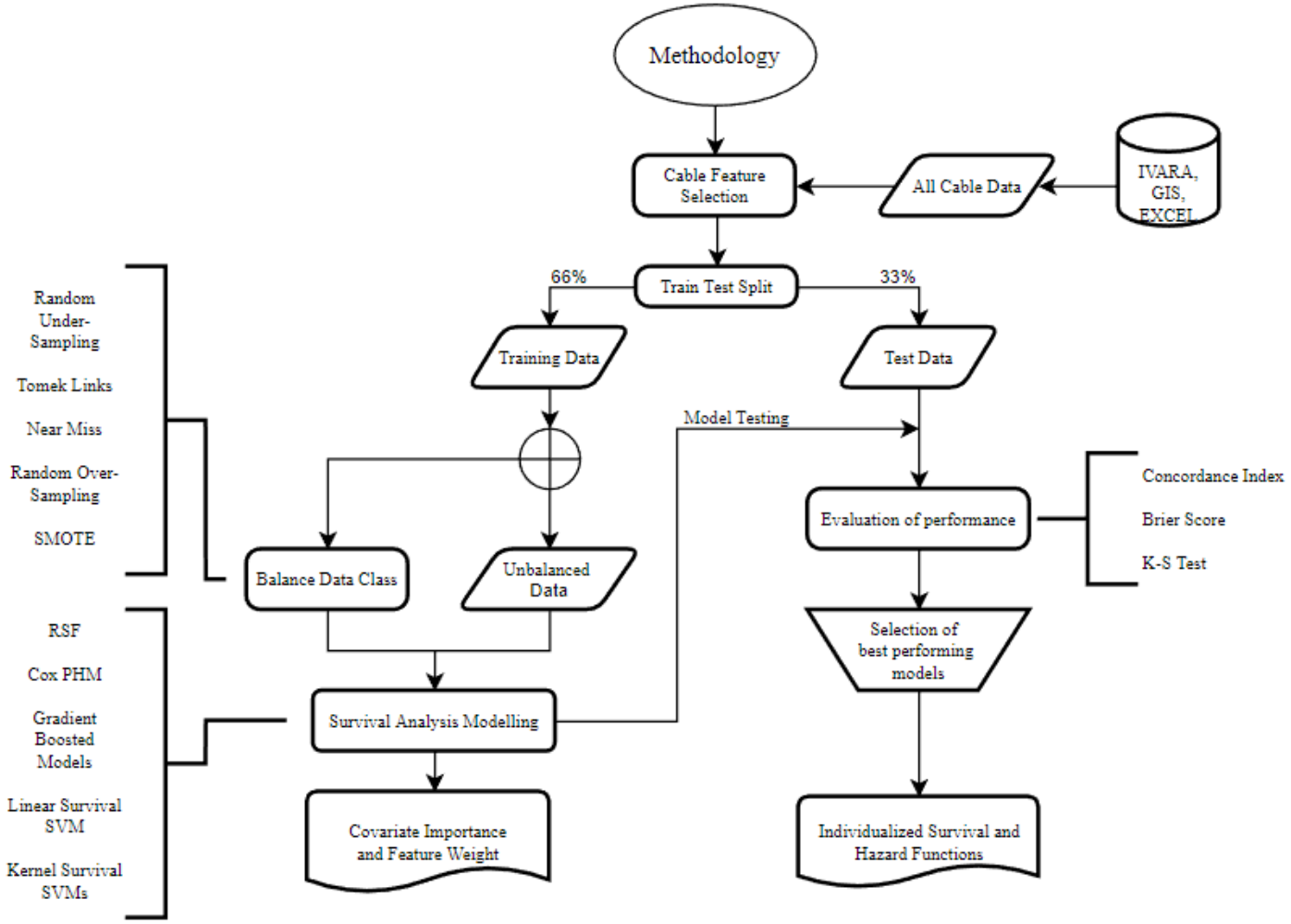


Figure 3.1: Methodology flowchart

3.1 Hypotheses

Four hypotheses address the objective of conducting a comparative assessment of various survival analysis techniques and class balancing methods to achieve high predictive performance, capable of extending to individualized hazard and survival curves for underground cables:

1. cable properties have an observable and quantifiable effect on cable failure;
2. class balancing has a positive outcome on the performance results of survival models;
3. the performance of the models is comparable with the same metrics; and
4. the model can be related to cables on an individual basis.

Each subsequent hypothesis relies on the validation of the previous hypothesis. By providing experimental support for the hypotheses, it becomes possible to establish a methodology to generate an accurate model capable of highlighting the risk factors and their effects on underground cable on a grouped and an individual basis.

3.1.1 Comparative Assessment Requirements

A comparative assessment of modelling methods must have the following characteristics: the effects of the covariates in the study have an observable effect on cable hazard; the use of class balancing provides additional insight into the data used and the performance of survival models and is comparable using performance metrics; and the model can be extended to cables on an individualized basis.

A successful assessment of the models had to meet all four criteria, and so it was necessary to validate all the hypotheses to meet the objectives of the study. The first criterion depended on the type and quality of the survival data used: either the data was able to provide insight into risk factors generated by the models or it did not. The second criterion defined methods of dealing with highly censored data in a useful manner so as to improve the performance of the survival models. The third criterion established a method for comparing the

performance of various survival models to determine how they perform relative to one another. The fourth criterion allows for cables to be evaluated in terms of their survival probability on an individual basis with respect to their specific parameters.

3.2 Feature Extraction

A desired characteristic of the survival analysis method applied was that little *a priori* cable knowledge was required to build a survival model that was capable of analyzing all of the data inputted. Minimizing the amount of data sorting and selective management of covariates allowed the models to train on more information and determine the covariates' importance without the requirement of covariate pre-selection.

3.2.1 Cable Characteristics

Medium-voltage (MV) distribution cable operates at voltages above 11 kV and below 50 kV. Many of these cables have been in operation for over 50 years, longer than the expected lifetime of cross-linked polyethylene (XLPE) cable [2]- a common cable used in the power service provider's distribution system for its temperature and abrasion resistant properties.

Underground cables feature a multilayer construction (Figure 3.2) designed for minimizing damage and increasing the longevity of their operability [37]. However, many failure mechanisms still exist. Breakdown of the cable insulation due to partial discharge caused by localized electrical fields around any defects in the insulating material cause serious damage to cables if left undetected [95].

Distribution cable conductors are made primarily from copper or aluminum, both of which provide excellent electrical conductivity. The shift to aluminum conductors is backed by the inexpensive nature of using aluminum when compared to using copper conductors; a cost difference of about three to five times [27]. The use of aluminum, however, comes with increased risk of conductor corrosion. Corrosion and oxidation become greater risk factors, introducing additional resistance, and resulting in additional, unwanted, heat generation in the

distribution system. Copper, on the other hand, is far more resistant to corrosion and other chemical exposure [91].

Another failure mechanism pertinent to the use of cables underground is the appearance of water-tree induced faults. Water-treeing occurs in cables that have been exposed to high moisture levels wherein defects occur in the insulation and semiconducting materials [119]. This phenomenon quickly degrades the cable by creating localized stresses at the point of the water tree.

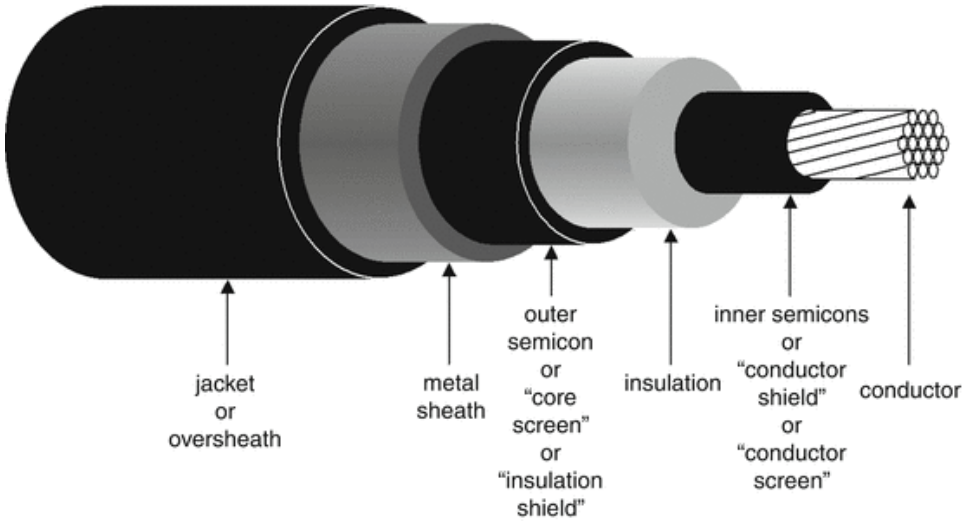


Figure 3.2: Cross-linked polyethylene cable cross section [129]

3.2.2 Data Selection

Data used in the study was selected on the basis of the availability of information pertaining to covariates and their completeness and based on the failure mechanisms discussed in Section 3.2.1.

A dataset of 8,172 cable instances was created, all of which contained complete covariate information. In order to add a greater number of entries, estimation of *interval* parameters would be required using the mean covariate values. In doing so, the models would become reliant on the use of estimated parameters rather than true value measurement and would further complicate the system when using class balancing.

Seven covariates of interest were included, with the addition of cable age and failure status, to complete the survival data triplet. The covariates include cable diameter, length, insulating and conducting material, number of cables in the configuration, number of repair splices, and the arrangement in which the cables are buried underground. The covariates and their respective scale of measure are summarized in Table 3.1.

Table 3.1: Covariates and scale of measure

Covariate	**Scale of Measure**
Cable Age	Interval
Cable Diameter	Interval
Cable Length	Interval
Conducting Material	Nominal
Insulation Material	Nominal
No. of Cables	Interval
No. of Splices	Ratio
Underground Arrangement	Nominal
Failed Status	Binary

Three covariates, namely the conducting material, insulating material, and underground arrangement, are categorical in nature (Table 3.2). Categorical variables require additional pre-processing in order to extract information out of the data by transforming the qualitative properties into quantitative measures. *One-hot encoding* [15] creates a set of binary variable columns corresponding to the number of categories in a covariate to create binary encoded values that machine learning models can utilize as shown in Figure 3.3. The original 8,172 × 9 tuple is thus transformed to a 8,172 × 16 tuple using one-hot encoding.

Table 3.2: Instance count of categorical covariates

Categorical Covariate	Number of Instances
Conducting Material	*Aluminum*: 7494
	Copper: 678
Insulation Material	*EPR CN*: 76
	EPR LC Shield: 4
	PILC: 57
	TRXLPE: 41
	XLPE CN: 7992
	XLPE LC Shield: 2
Underground Arrangement	*Directly Buried*: 8098
	Duct line: 74

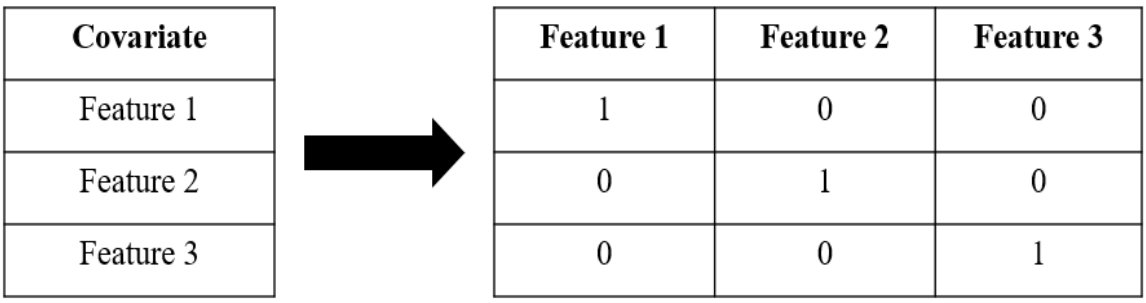


Figure 3.3: One-hot encoding for categorical covariates

3.3 Implementing Class Balancing Techniques

Class balancing is pertinent to the success of the study. Class balancing can be performed on any covariate, in attempt to balance the number of feature instances. In this study, class balancing is performed on the failure status of the cable as the number of censored cases far outweighs the number of uncensored cases; this is to be expected when using real-world survival data.

The objective of class balancing allows machine learning models to be trained with minimal bias. A heavily imbalanced dataset creates bias in machine learning models, skewing the model's predictive capability. The resultant is an accuracy paradox where overly accurate results are generated but are not indicative of the true class prediction of the data. The predictions made by the model in a highly imbalanced data tend to overestimate the number of predicted majority class instances because the model itself is trained on imbalanced classes. In doing so, the minority class is largely outweighed, and the number of false positive and false negative results increases.

The class balancing techniques used in this study are adopted from the Python package, *imbalanced-learn* (*imblearn*) [82].

Random Under-Sampling

In the case of random under-sampling (RUS), data instances from the majority class are removed naively. That is, instances are removed in a completely random manner without replacement, as represented in Figure 3.4. Although simple and effective, the limitation of RUS is that this method does not consider the usefulness and importance of the information that is removed. It is likely that in a heavily imbalanced dataset, RUS eliminates useful information from the dataset when balancing the classes.

The usability of RUS is straightforward in procedure and is implemented by:

1. Setting a minority, $\boldsymbol{m}$, and majority class, $\boldsymbol{M}$,
2. selecting an instance, x_i, at random, from $\boldsymbol{M}$, and eliminating it, and
3. repeating step 2 until the number of instances from $\boldsymbol{m}$ and $\boldsymbol{M}$ is equivalent

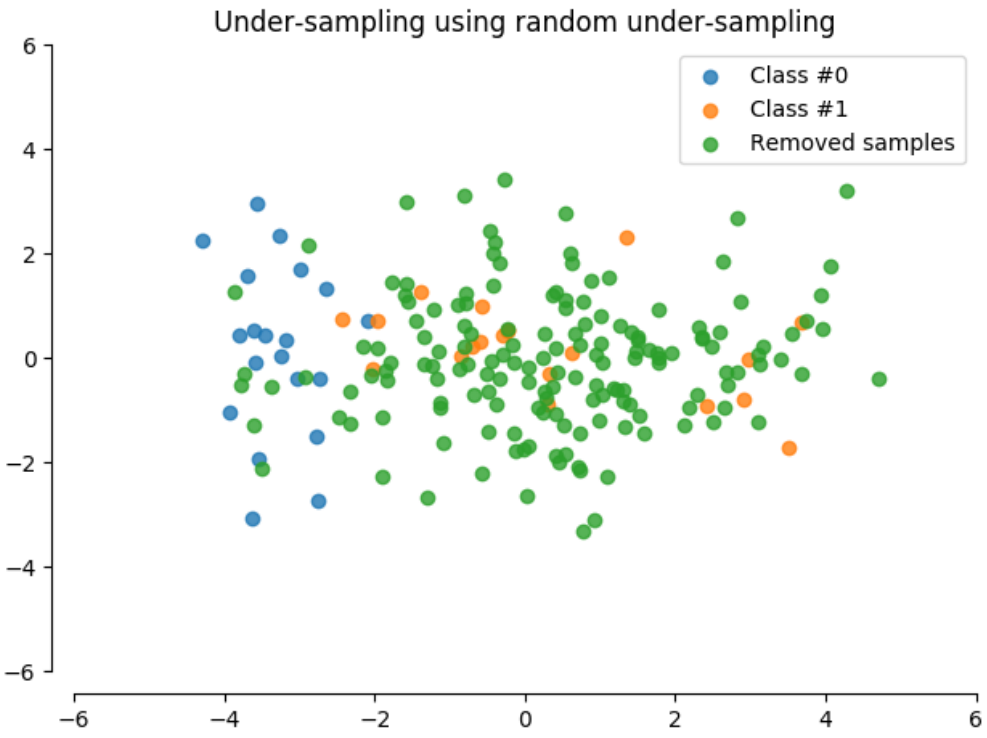


Figure 3.4: Representation of random under-sampling balancing [81]

Tomek Links

Tomek Links is formulated as a modified condensed nearest-neighbours sampling method. A so-called Tomek Link (Figure 3.5) is defined as points from the majority class, x_M, and minority class, x_m, that are distance, $d(x_M, x_m)$, apart provided that no other class, x_z, such that $d(x_m, x_z) < d(x_M, x_m)$, and $d(x_M, x_z) < d(x_M, x_m)$, exists.

When Tomek Links have been identified in the data feature space, the majority class instances are eliminated by applying the nearest-neighbour rule to select the instances [40].

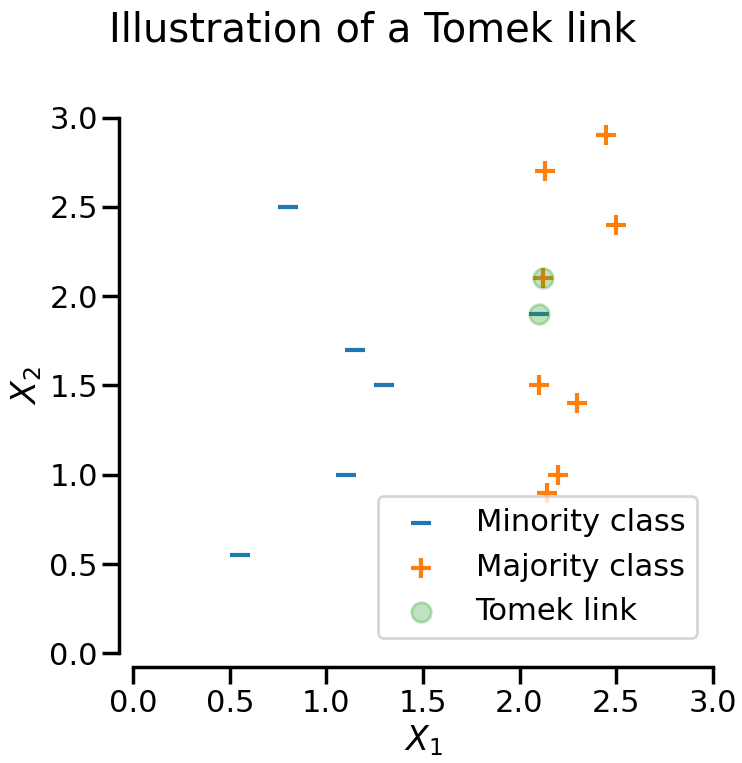


Figure 3.5: Representation of Tomek Links balancing [81]

NearMiss

NearMiss is used on two instances of different classes, in close proximity to one another in the feature space, by removing that of the majority class in order to increase the dispersion of points in the classes. This is a commonly used under-sampling technique as it provides the added benefit of preserving information when compared to RUS and other under-sampling techniques.

NearMiss can be applied to a dataset by:

1. Setting a minority, $\boldsymbol{m}$, and majority class, $\boldsymbol{M}$,
2. calculating the Euclidean distance between all instances in $\boldsymbol{m}$ against $\boldsymbol{M}$,
3. setting an under-sampling rate, N, with the number of instances that, when removed from the majority class, will balance the $\boldsymbol{m}$ and $\boldsymbol{M}$ instances, and
4. selecting N instances from $\boldsymbol{M}$ for which average Euclidean distance is smallest to instances in $\boldsymbol{m}$ and eliminating them.

Random Over-Sampling

Random over-sampling (ROS) enables the preservation of critical information in the dataset, unlike RUS. ROS involves generating new instances in the minority class by duplicating instances randomly. ROS is effective for machine learning algorithms where distributions are skewed and where duplicate instances can influence the fit of a model in a positive manner. However, this comes with an increased computational cost and increased likelihood of overfitting.

Applying ROS to a dataset involves:

1. Setting a minority, $\boldsymbol{m}$, and majority class, $\boldsymbol{M}$,
2. selecting an instance, x_i, at random, from $\boldsymbol{m}$, and duplicating it, and
3. repeating step 2 until the number of instances from $\boldsymbol{m}$ and $\boldsymbol{M}$ is equivalent

SMOTE

SMOTE synthesizes minority class data between existing minority instances through linear interpolation of data. The new data points are generated by randomly selecting k-nearest-neighbours and synthesizing data in the feature space. Satphathy [114] provides a graphical representation of the SMOTE algorithm (Figure 3.6) where r_1 is the synthetic point generated out of two k-nearest-neighbour points, X_1 and X_{11}.

Algorithmically, SMOTE can be performed on a dataset by:

1. Setting a minority class, $\boldsymbol{m}$,
2. for each point x, calculating the Euclidean distance between all $x \in \boldsymbol{m}$ to determine the k-nearest-neighbours of x,
3. setting a sampling rate, N, to create balanced classes,
4. randomly select $x \in \boldsymbol{m}, N$, i.e., $x_1, x_2, \ldots x_N$, from k-nearest-neighbours and create a new dataset, $\boldsymbol{m_1}$, and
5. for each $x_i \in \boldsymbol{m_1}$, $i = 1, 2, \ldots, N$ generate synthetic data using:

$$\mathrm{x}' = \mathrm{x} + \mathrm{rand}(0,1)|\mathrm{x} - \mathrm{x_i}| \tag{3.1}$$

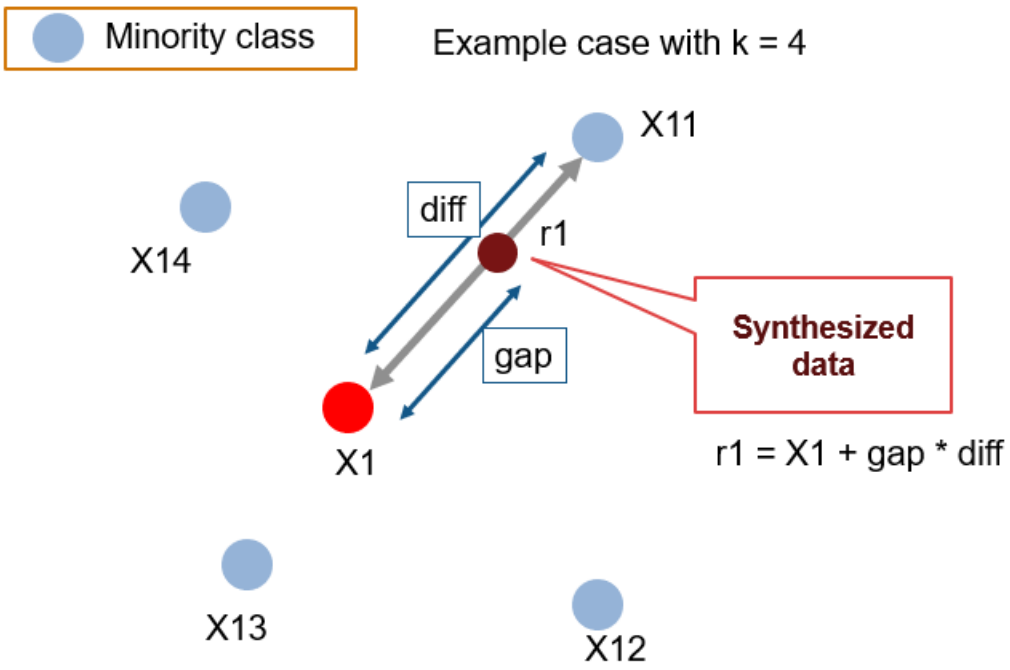


Figure 3.6: Representation of SMOTE balancing [114]

3.4 Modelling for Survival Analysis

Once class balancing techniques have been performed on the data, survival models can be utilized to further the understanding between the covariates of interest and the likelihood of survival of cable instances.

The survival models used in this study are selected based on the availability of documentation related to their applicability in survival analysis, general performance and feature identification ability, and satisfaction of the hypotheses in the study.

Cox Proportional Hazards Model

Cox PHM is among the most widely adopted models for survival analysis. The *hazard rate,* or the risk of failure given that a data instance has survived up to a time, t, is the primary measure of effect when using Cox PHM.

Cox PHM is given by,

$$h(t, \boldsymbol{x}) = h_0(t) \exp(\sum_{i=1}^{n} \beta \boldsymbol{x_i}) \quad (3.2)$$

where the expected hazard function, $h(t, \boldsymbol{x})$, at a time, t, with covariates, $\boldsymbol{x}$, is derived from the baseline hazard, $h_0(t)$, covariate coefficients, β, and covariate, $\boldsymbol{x_i}$. A key feature in Cox PHM is that $h_0(t)$ is derived from the model itself, calculated as the hazard function if all $\boldsymbol{x_i} = 0$, and requires no estimation of the parameter which contributes to increased model performance.

The Cox PHM model relies on several assumptions that are critical to its use in survival analysis:

1. All instances in the data as presumed to be independent, i.e., the survival time of one instance does not inform the estimated survival of another instance,
2. censoring of data is done in a non-informative or independent manner; the assumption is satisfied when there is no relationship between failure and probability of censoring,
3. the relationship between the log hazard and each covariate is linear, and
4. the proportional hazard assumption- the assumption that all instances have the same hazard function but possess unique scaling factor specific to the covariates of the instance. This implies that the effect of a risk factor is constant over time.

Gradient Boosted Models

Gradient boosted models perform better than traditional survival analysis models in the case of high-dimensional data, where the analysis is computationally expensive and variable selection becomes increasingly complex. Rather than fitting base-learners to the entire covariate feature set, component-wise gradient boosting fits covariates successively by selecting the best covariate iteratively [21].

Two gradient boosted models are used- a gradient boosted model that implements regression tree base learners which is efficient in determining non-linear relationships between covariates and survival time, and a model that encompasses component-wise least squares as the base learner.

The overall gradient boosted model, GB, is an additive model consisting of multiple base learners,

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} \beta_m g(\boldsymbol{x}; \theta_m) \tag{3.3}$$

where $M > 0$ is the number of base learners, $\beta_m \in \mathbb{R}$ is the weight function associated with the model, and the function g is the base learner that is parameterized by the parameter, θ. In regression gradient boosted models, specification of a loss function, $\phi(\boldsymbol{y}, f(\boldsymbol{x}, \theta))$ is pertinent. The loss function must be a convex function [98] that can be drawn from the negative partial log-likelihood function based on Cox PHM [104],

$$\phi(\delta, f(\boldsymbol{x})) = \arg\min_{f} \sum_{i=1}^{n} \delta_i \left[f(\boldsymbol{x_i}) - \log\left(\sum_{j \in R_i} exp\left(f(\boldsymbol{x}_j) \right) \right) \right] \tag{3.4}$$

To implement gradient boosting models in survival analysis,

1. choose an initial $f_0(\boldsymbol{x})$**,** typically a value of zero,
2. add a new base learner function at each iteration, keeping parameters and coefficients constant,
3. find the base learner, $g(\boldsymbol{x}; \theta)$, and its respective weight β_m, such that $\phi(\delta, f(\boldsymbol{x}))$ is minimized,
4. add the result, $\beta_m g(\boldsymbol{x}; \theta)$, from Step 3 to the model to create a new model based on the $(m-1)^{\text{th}}$ iteration, and
5. repeat Steps 2 – 4 until $\phi(\delta, f(\boldsymbol{x}))$ cannot be minimized further or until iteration, m, equals the stop iteration point, M.

Random Survival Forests

Random survival forests have proven to be a favorable model when the proportional hazard assumption of Cox PHM is violated. Randomization in RSF is two-fold [66]. Firstly, a randomly drawn bootstrap sample of the data is used to grow the trees of the forest. Secondly, the trees depths are increased by using splitting nodes with randomly chosen covariates.

The RSF algorithm is described by Ishwaran et al. [66] as follows:

1. Draw n tree bootstrap samples from the original dataset using approximately 67% of the data (the remaining data is considered out-of-bag data),
2. grow a survival tree for each bootstrapped sample where, at each node of the tree, covariates are randomly selected for testing the split,
3. split on a covariate using the *log-rank splitting rule* where the covariate maximizes the survival differences across the daughter nodes,
4. grow the trees to maximum depth, i.e., as close to saturation as possible where each terminal node has no less than the node size, $d_0 > 0$, events,
5. calculate the ensemble cumulative hazard function (CHF) estimate by combining the CHF from all trees, and
6. using the out-of-bag data, calculate prediction error for the ensemble CHF.

The log-rank splitting rule is the default splitting rule used as it provides the highest model accuracy when compared to the *conservation of events, log-rank score,* and *log-rank approximation* splitting rules [44, 67]. The log-rank split at the value, c, for covariate, $\boldsymbol{x}$, is,

$$L(x,c) = \frac{\sum_{i=1}^{N}\left(d_{i,1} - Y_{i,1}\frac{d_i}{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\frac{Y_{i,1}}{Y_i}\left(1-\frac{Y_{i,1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i}} \tag{3.5}$$

Where $d_{i,j}$ is the number of deaths at time, t, and $Y_{i,j}$ is the number of individuals in daughter node j, who are alive at or have an event at time t_i. Y_i and d_i are defined as,

$$Y_i = Y_{i,1} + Y_{i,2} \tag{3.6}$$

$$Y_{i,1} = \#\{T_l \geq t_i, x_l \leq c\}, \quad Y_{i,2} = \#\{T_l \geq t_i, x_l > c\} \tag{3.7}$$

and,

$$d_i = d_{i,1} + d_{i,2} \tag{3.8}$$

where T_l is defined as the event time for individual l, in the case that the cable has failed, or right censored time, in the case that the cable has not failed at the end of the study. $|L(x,c)|$ is the measure of node separation; the larger the value, the better the split is. The best split at a node is

found from the covariate, x^*, and split value, c^*, where $|L(x^*, c^*)| \geq |L(x, c)|$ for all x and c [67].

The CHF is found for each terminal node, h, using the Nelson-Aalen estimator,

$$\widehat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \tag{3.9}$$

where all cases in h have the same CHF [66]. The out-of-bag ensemble CHF for instance, i, is then calculated over the average of B survival trees by,

$$H_e^{**}(t|\boldsymbol{x}_i) = \frac{\sum_{n=1}^{B} I_{i,n} \widehat{H}_h(t)}{\sum_{n=1}^{B} I_{i,n}} \tag{3.10}$$

where $I_{i,n} = 1$ if instance i is an out-of-bag case for the n^{th} bootstrap sample, otherwise $I_{i,n} = 0$.

Survival Support Vector Machine

Survival analysis can be cast as a learning-to-rank problem; cables with a lower predicted survival time should be ranked before cables with longer survival time. However, in the case of censored data, which is highly pertinent to the analysis, pairwise samples used in model training that are both censored, $\delta_i = \delta_j = 0$, it is unclear whether the i-th sample is to be ranked before or after the j-th sample as the time to failure is unknown. The same applies to the pairwise comparison of one uncensored sample and one censored sample ($\delta_i = 1$ and $\delta_j = 0$). Thus, in model training, the set of valid pairwise comparable instances, P, is given by [105],

$$\mathrm{P} = \{(\mathrm{i}, \mathrm{j}) \mid \mathrm{y_i} > \mathrm{y_j}, \delta_\mathrm{j} = 1\}_{\mathrm{i,j=1}}^{\mathrm{n}} \tag{3.11}$$

Training of a linear survival SVM requires solving and minimizing the loss function through Equation 3.12:

$$\arg\min_{\mathbf{w},\mathrm{b}} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \cdots$$
$$\frac{\alpha}{2}\left[\mathrm{r}\sum_{\mathrm{i,j}\in\mathrm{P}} \max\left(0, 1 - \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_\mathrm{i} - \mathbf{w}^{\mathrm{T}}\mathbf{x}_\mathrm{j}\right)\right)^2 + (1-\mathrm{r})\sum_{\mathrm{i}=0}^{\mathrm{n}}\left(\zeta_{\mathbf{w},\mathrm{b}}(\mathrm{y_i}, \mathbf{x}_\mathrm{i}, \delta_\mathrm{i})\right)^2\right] \tag{3.12}$$

where,

$$\zeta_{\boldsymbol{w},b}(y_i, \boldsymbol{x}_i, \delta_i) = \begin{cases} \max(0, y_i - \boldsymbol{w}^T \boldsymbol{x}_i - b) & \text{if } \delta_i = 0, \\ y_i - \boldsymbol{w}^T \boldsymbol{x}_i - b & \text{if } \delta_i = 1 \end{cases} \tag{3.13}$$

The model coefficient is defined as $\boldsymbol{w} \in \mathbb{R}^{\mathrm{p}}$, d-dimensional covariate vector, $\boldsymbol{x}_i$, and the survival time or time of censoring, $y_i > 0$. Of note is the hyper-parameter, $\alpha > 0$, which determines the degree to which regularization is applied and the hyper-parameter $r \in [0,1]$ which reduces the model to a ranking objective if $r = 1$ or a regression objective if $r = 0$.

Kernel-based survival SVMs are a generalization of linear survival SVMs that can solve more complex data through the use of non-linear functions. Pölsterl et al. [105] describe this process as reformulating Equation 3.12 with respect to finding a function $f: \chi \to \mathbb{R}$ to a kernel function (Equation 3.14): $k: \chi \times \chi \to \mathbb{R}$. Kernel survival SVMs, while more complex, use the kernel trick to represent the data through pairwise similarity comparisons between data instead of explicitly applying a transformation as would be done with a linear survival SVM.

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{3.14}$$

Various kernel functions can be employed in the kernel survival SVM, which may be disadvantageous to the selection of the most appropriate kernel for the application and for tuning hyper-parameters. All kernels shown in Table 3.3 are employed and tuned to provide the most accurate results for each model.

Table 3.3: Kernel functions for survival SVM

Kernel	**Function**
Linear	$K(x_i, x_j) = x_i^T x_j$
Polynomial	$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \ \gamma > 0$
RBF	$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \ \gamma > 0$
Sigmoid	$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

Two-Stage Cox PHM and Random Survival Forest Model

Both Cox PHM and RSF possess desirable features fit for survival analysis. Combining the models in a two-step approach has proven to be a robust model capable of generate highly accurate results [11]. In this approach, high-dimensional data is evaluated using Cox PHM to determine the covariates of most significance. This is done by evaluating $|\beta|$ values of each covariate and ranking them from high absolute value to lowest. The higher the value, the larger the influence of the covariate on the model's hazard, either in a negative or positive manner. By selecting the top n covariates, RSF can be used to train and test the model's performance. The RSF stage is identical to that of the conventional RSF model.

3.5 Performance Metrics

Evaluating the performance of the models using comparable metrics is vital to understanding the capabilities of each model, as well as determining which model is best suited for analyzing the cable dataset presented. This is the basis of the third hypothesis.

Concordance Index

The C-index is designed to estimate the concordance probability $\mathrm{P}(\eta_j > \eta_i \mid T_i > T_j)$ such that two independent instances are ranked based on risk scores, η and event or censoring time, T. The C-index is a generalization of the area under the ROC curve that accounts for censored data and represents the discriminatory power of a model. A model with perfect predictive accuracy is given a C-index value of 1, a model that does as good as random guessing has a C-index value of 0.5, and a model that has no predictive capability has a value of 0.

The C-index of the model is computed for every pair of cable instances, i and j, $i \neq j$, by evaluating η and T.

1. If $\delta_i = \delta_j = 1$, i.e., not censored, the pair (i, j) is a *concordant pair* if $\eta_i > \eta_j$ and $T_i < T_j$, or a *discordant pair* if $\eta_i > \eta_j$ and $T_i > T_j$
2. If $\delta_i = \delta_j = 0$, i.e., censored, it is unclear which cable failed first and the pair is omitted from the computation

3. If, for example, $\delta_i = 1$ and $\delta_j = 0$, i.e., one cable is censored and the other is not and,
    a. $T_i > T_j$, the pair is omitted from the computation
    b. $T_i < T_j$, cable i failed first and the pair (i, j) is a *concordant pair* if $\eta_i > \eta_j$, or a *discordant pair* if $\eta_i < \eta_j$.
4. Compute the C-index as:

$$\mathrm{C-index} = \frac{\#\ \text{of concordant pairs}}{\#\ \text{of concordant pairs} + \#\ \text{of discordant pairs}}$$

The above can be alternatively expressed by the formula [58, 116]:

$$\mathrm{C-index} = \frac{\sum_{i,j} I(T_i > T_j) \cdot I(\eta_j > \eta_i) \cdot \delta_j}{\sum_{i,j} I(T_i > T_j) \cdot \delta_i} \tag{3.15}$$

Brier Score

The Brier score is used as a performance metric for survival problems that incorporate censored data. It is defined as a measure of the square deviation of the survival estimate from the true probability of failure.

The probability distribution of being non-censored until time t, is given by the function $G(t) = P(T > t)$ and the estimate of $G(t)$ is denoted as $\hat{G}(t)$. Incorporating the estimated survival function, $\hat{S}(t|x_i)$, the Brier score is described by,

$$BS(t) = \frac{1}{N} \sum_{i=1,\dots,N} \begin{cases} \frac{\left(0 - \hat{S}(t|x_i)\right)^2}{\hat{G}(t_i)} & \text{if } t_i \le t, \delta = 1 \\ \frac{\left(1 - \hat{S}(t|x_i)\right)^2}{\hat{G}(t)} & \text{if } t_i > t \\ 0 & \text{if } t_i = t, \delta_i = 0 \end{cases} \tag{3.16}$$

The Brier score evaluates the model's goodness of fit for a given time, t, and is extended to be an overall measure for the model's prediction at all times- the integrated Brier score (IBS):

$$IBS = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t) \mathrm{d}t \tag{3.17}$$

3.5.1 Underlying Distribution Analysis

The K-S test is employed to determine whether the empirical CDF generated by the model comes from a population with a specific distribution. The empirical CDF, by definition, is obtained through the survival function,

$$S(t) = \exp\left(-\mathrm{H}(t)\right) \tag{3.18}$$

where the hazard function, $\mathrm{H}(t)$, is calculated directly from the model's output. The empirical CDF, $F(t)$, by definition, is then solved through:

$$F(t) = P(T \leq t) = 1 - S(t) \tag{3.19}$$

The K-S test, while a robust distribution analysis tool, has several limitations:

1. It has an increased sensitivity near the center of the distribution than at the tails,
2. it only applies to continuous distributions, and
3. the distribution must be fully specified, namely the location, shape, and scale parameters

Two outputs are the resultant of performing a K-S test- the test statistic, D, and the significance, p. The former is the absolute maximum supremum between the CDFs, defined as,

$$D = \max_{1 \leq i \leq N} \left(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right) \tag{3.20}$$

where $Y_i = Y_1, \ldots, Y_N$ are N ordered data points. The closer the value of D to zero, the more likely that the empirical CDF was drawn from the same distribution. More specifically, pre-determined critical values, α, determine whether the null hypothesis is rejected. For larger samples, the null hypothesis is rejected if,

$$D_{n,m} > \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1+\frac{m}{n}}{2m}} \tag{3.21}$$

where, m and n are sample sizes for the two distributions. The latter is the conventional interpretation of the p-value. That is, the measurement used to validate a hypothesis. The p-value itself, is evidence against a null hypothesis; the smaller the p-value, the stronger the evidence is supporting the rejection of the null hypothesis.

3.6 Summary of Methodology

It is recognized that survival analysis follows a sequential approach to achieve the desired objective- a highly accurate model capable of strong predictive performance. Since the level of imbalance is known *a priori,* additional steps to further the performance of the models are added; the underlying limitation in survival analysis is mitigated using class balancing techniques. Various survival modelling techniques are employed as models of interest, based on prior performance in conventional survival analysis applications.

The methodology outlined in Figure 3.1 provides a novel approach to conducting a survival analysis to generate outputs from various combinations of class balancing techniques and survival models. The four-stage approach to modelling ensemble hazard and survival functions for cable instances is concluded with an approach to individualizing these functions for specific cable instances.

This method builds off previous steps to create a model that outperforms conventional survival analysis techniques through the means of additional pre-processing measures and is validated through the comparative assessment of performance metrics for each model.

The next chapter examines the implementation and results of the four-stage approach for cable survival data.

Chapter 4

Experimental Results

Chapter Three introduced a methodological approach to conducting a survival analysis of medium-voltage underground cables with the inclusion of class balancing techniques that aims to validate four hypotheses:

1. cable properties have an observable effect on cable failure;
2. class balancing has a positive outcome on the performance results of survival models;
3. the performance of the models is comparable with the same metrics; and
4. the model can be related to cables on an individual basis.

This chapter outlines the results obtained from class balancing techniques, survival modelling, and performance evaluation for the medium-voltage underground cable dataset. Relationships between the choice of class balancing technique and survival model are drawn in an identifiable way through the use of performance metrics. Covariates of importance are examined and their respective influence on the hazard of specific cable instances are determined. Validation of the four hypotheses is simultaneously confirmed.

4.1 Feature Identification

Pair plots are created to identify statistical relationships between numerical variables including cable age, diameter, length, number of splices in the cable, and the number of cables in the configuration.

From Figure 4.1, an apparent relationship exists between phase configuration and age. Cable failure is more commonly identified in a range of cable ages that are in a one-phase and three-phase configuration, whereas cables in a two-phase configuration do not experience failure

as commonly; the latter can be contributed to the minimal use of two-phase cable in the distribution network. Two-phase cable is only used wherever one-phase cable has diverged. Three-phase cable is used at substations and in commercial and industrial applications. One-phase cable is abundantly used for most households and consumer level applications. The use of two-phase is, therefore, only required in areas in which customers are not in need of three-phase power and where one-phase power would not be sufficient.

Cables that contain more repair splices are less likely to experience failure because splices provide a cable with added longevity of operation by eliminating a damaged portion of cable with a new segment rather than replacement of the entire cable.

Increasing geometric length of a cable shows a negative relationship with cable age; shorter cable (≥ 500 m) instances possess a larger ratio of active cables when compared with larger cables (< 500 m).

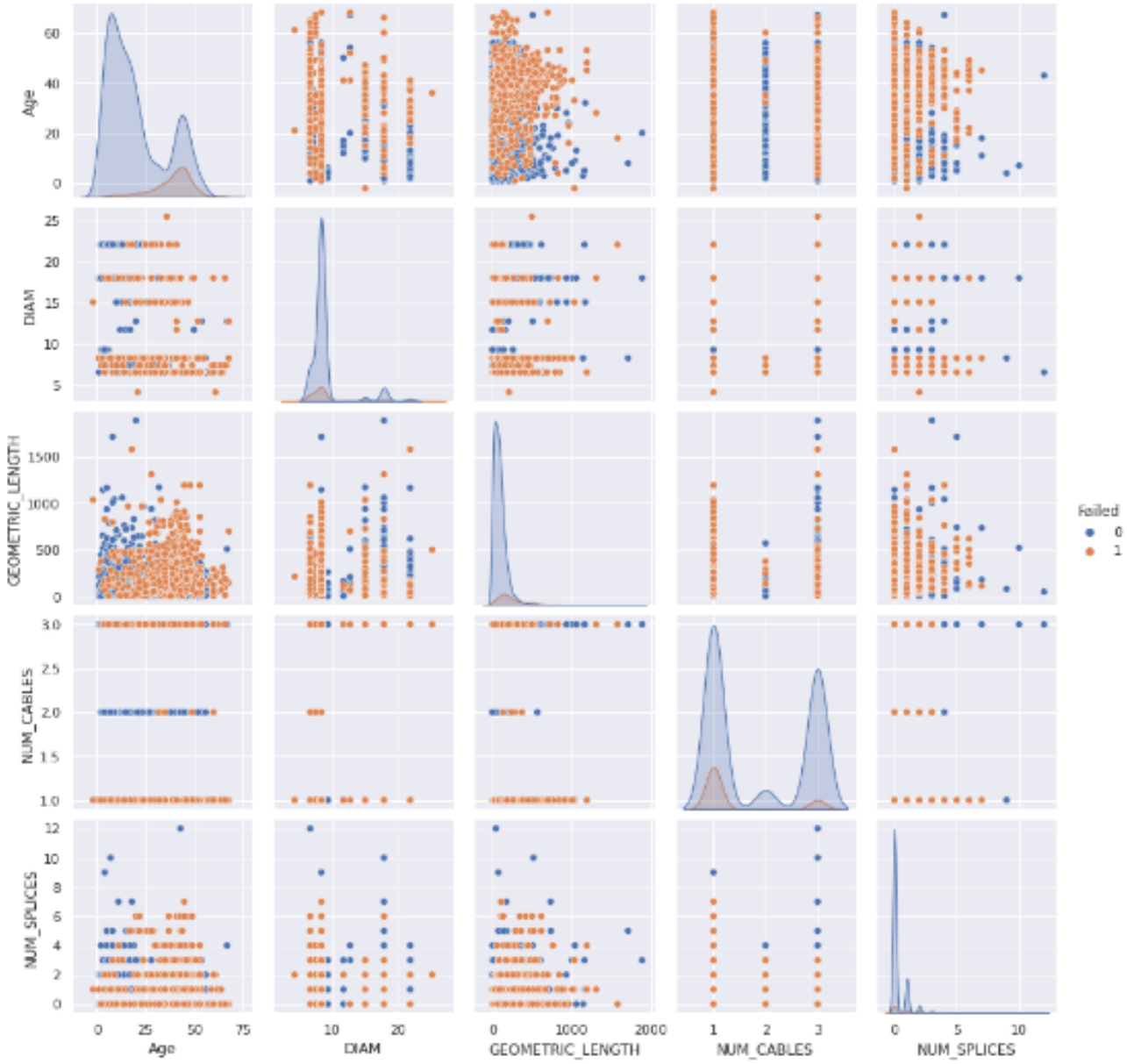


Figure 4.1: Seaborn pair plot of numerical covariates

Categorical data are then investigated to gauge failure instances. Categorical data includes insulating material, conducting material, and burial arrangement. The conducting material is predominately aluminum in cables used by the utility provider, primarily due to it being a cost-efficient alternative to previously used copper conducting material, hence the larger proportion of aluminum-based cable compared to older, less frequently used, copper-based cable (Figure 4.2).

Figure 4.3 provides information pertaining to the number of instances of various insulating materials of cables in the dataset. Cross-linked polyethylene (XLPE) is the most common type of insulating material used for underground cable in the region. XLPE is extremely resistant to abrasion and general wear; it also provides resistance to elevated voltage, chemical contamination, and water ingress. XLPE is most commonly used with concentric neutrals (XLPE CN) when compared with longitudinal corrugated XLPE (XLPE LC Shield) and tree retardant XLPE (TRXLPE).

Cables installed in a directly buried (DB) fashion are a more cost-efficient method of installation when compared with cable installed in a duct line conduit. Duct lines are used for specialty applications such as under roadways and from substations. Duct lines protect distribution cable from corrosion, temperature extremes, and seismic activity. Cables in duct lines are far more costly, with additional costs associated with manufacturing and installation duct lines prior to installing cable. Cable laid directly in the soil without a manufactured surrounding medium can be used in most situations. Due to the protective nature of cable insulation, cable that is directly buried still possesses resistance to the surrounding environment without the requirement for duct lines- the more common approach (Figure 4.4).

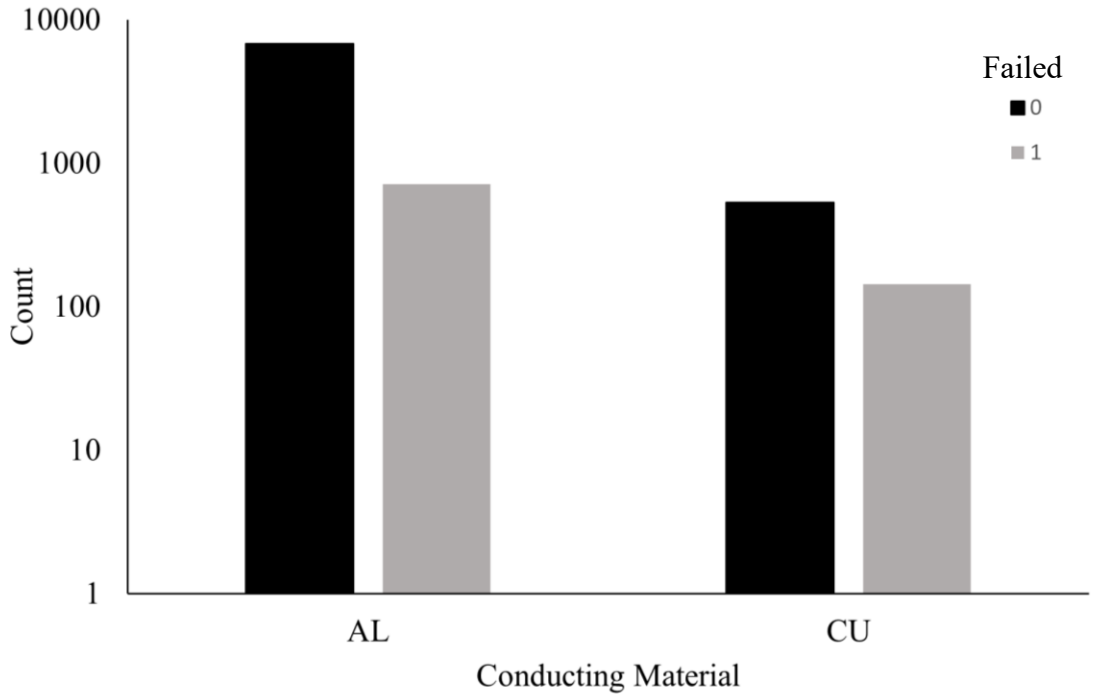


Figure 4.2: Conducting material instance count

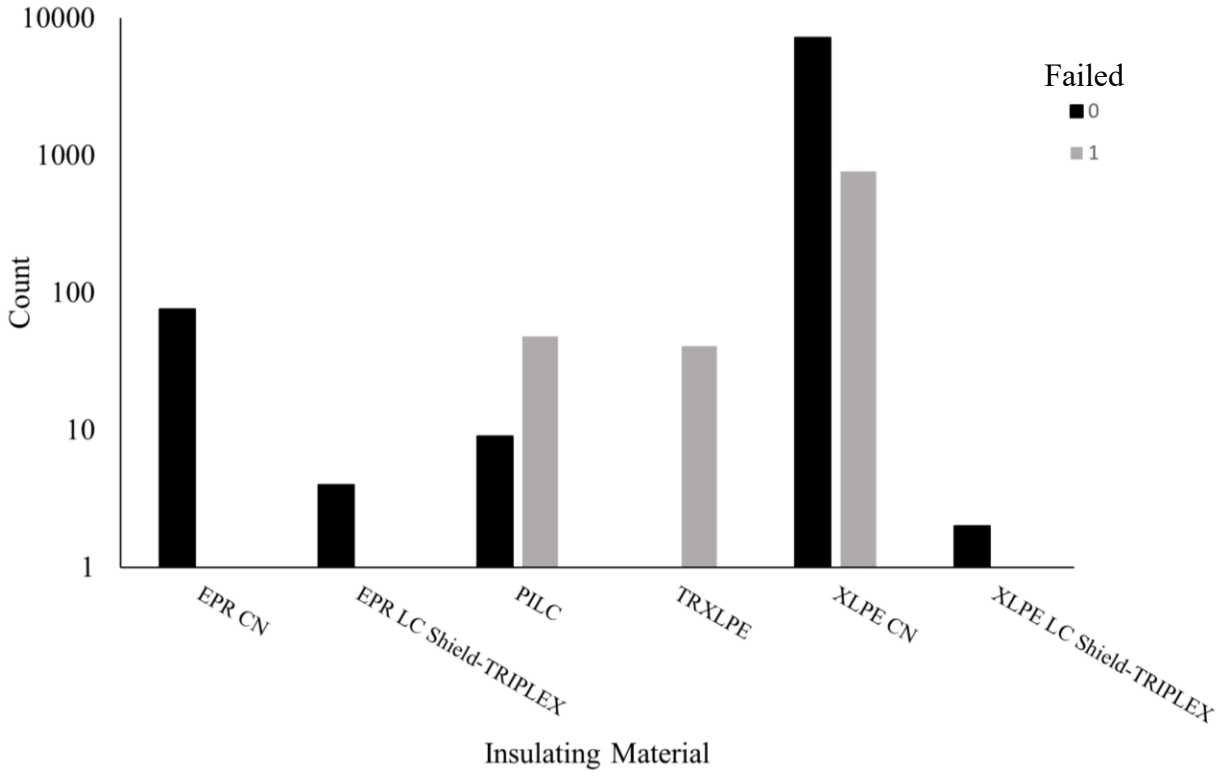


Figure 4.3: Insulating material instance count

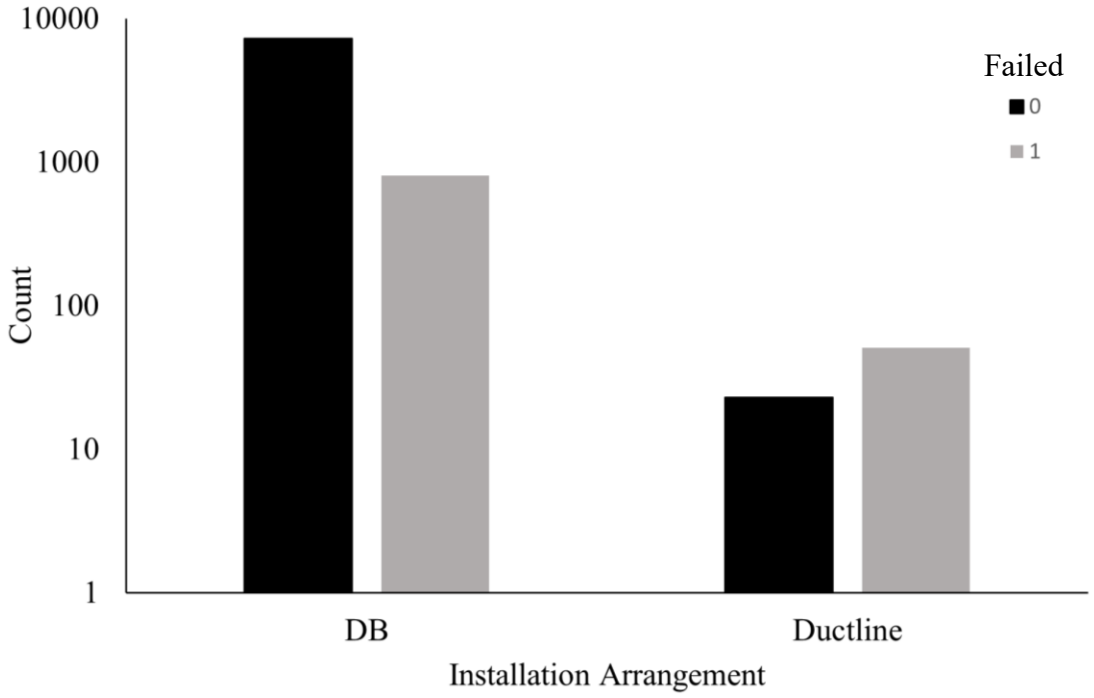


Figure 4.4: Installation arrangement instance count

Class Balancing

Class balancing refers to balancing unequal proportions of samples. Methods exist both for increasing the number of samples in the minority class to balance that of the majority class and for decreasing the number of samples in the majority class to balance with the minority class.

Due to the nature of real-world survival data, class balancing is conducted solely on the basis of failed ($\text{Class} = 1$) and active cable ($\text{Class} = 0$) status, with the proportion of active cable being far greater than that of failed cable in the dataset (Figure 4.5 (a)). Under-sampling techniques (Figure 4.5 (b)-(d)) remove instances from the majority class until the number of failed and active cable is equivalent. This, however, is not the case in Tomek Links, which, as discussed in Section 3.3, is a modification of the nearest-neighbour rule whereby instances of opposite classes that are nearest-neighbours are used to eliminate instances of the majority class. In doing so, Tomek Links do not provide a true class balance, but instead create a larger distance between classes in the feature space for a potentially more accurate analysis. Both over-sampling techniques (Figure 4.5 (e)-(f)) create an equal ratio of failed and active cable by adding instances to the minority class. In random over-sampling, random minority class instances are duplicated

to equate the minority and majority class instances. In the SMOTE technique, synthetic instances with slight variation to original instances are created in the minority class. This is beneficial for minimizing overfitting issues. Numerical results for the transformed data using class balancing techniques is summarized in Table 4.1.

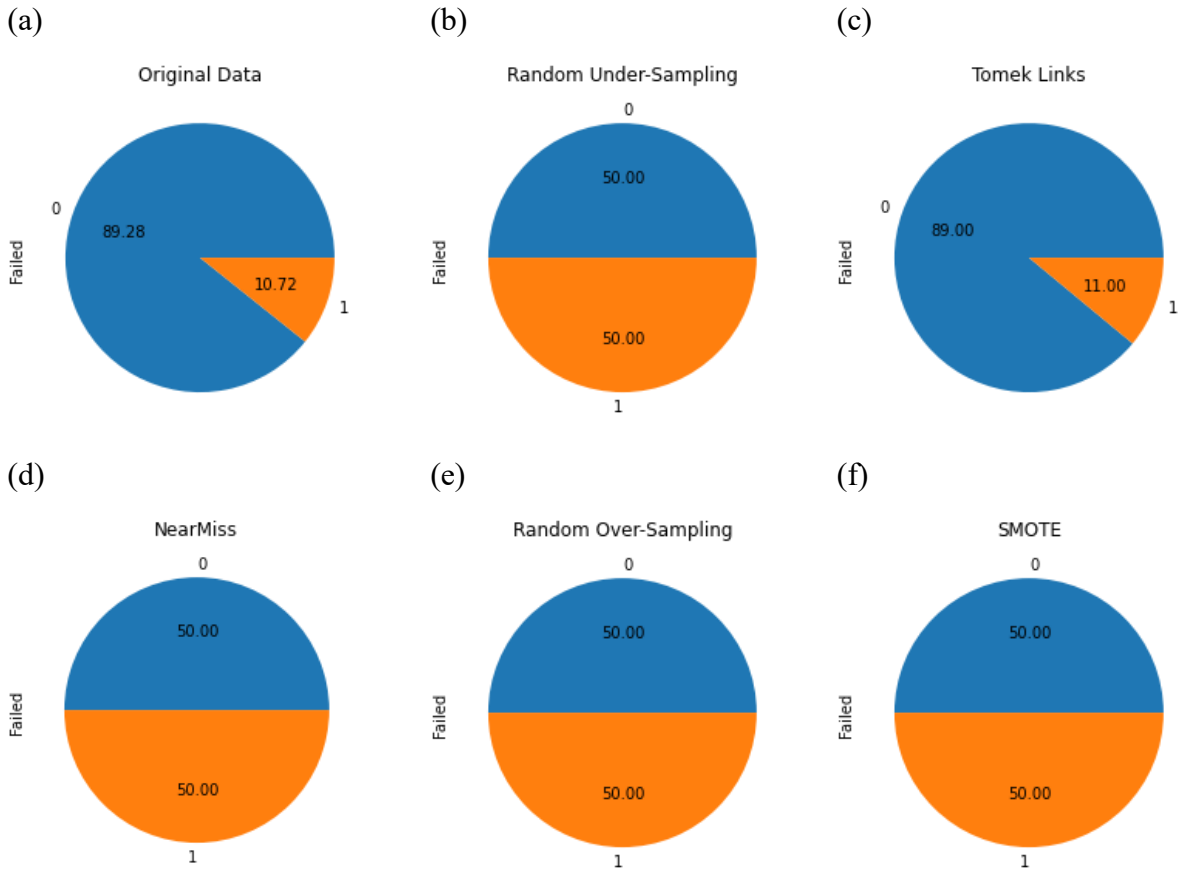


Figure 4.5: Ratio of censored and uncensored samples for (a) unbalanced data, (b)-(d) under-sampling methods, and (e)-(f) over-sampling methods

Table 4.1: Resampled data count using class balancing

Balancing Technique	**Original Training Data**	**Resampled Training Data**
Random Under-Sampling		*0:* 587
		1: 587
Tomek Links		*0:* 4747
		1: 587
NearMiss	*0:* 4888	*0:* 587
	1: 587	*1:* 587
Random Over-Sampling		*0:* 4888
		1: 4888
SMOTE		*0:* 4888
		1: 4888

In the pre-processing stage, the class balancing techniques are applied to the dataset after the train-test data split to ensure that duplicated instances in both under-sampling and over-sampling methods are not present in both the training and test data.

4.2 Survival Assessment

The data pre-processing stage leads into the construction and evaluation of various survival models using numerous performance metrics concurrently. The nine survival models used with the five class balancing techniques are compared against the unbalanced (original) dataset as a benchmark.

4.2.1 Optimization of Hyperparameters

Models that require additional parameterization of the number of estimators used include RSF, and gradient boosted models using both regression trees and least squares base learners. The number of estimators used in the model has a direct correlation to the C-index output from fitting the test data.

To determine the number of estimators that provides the largest C-index value, the number of estimators, i.e., the number of trees, in RSF is varied from 10 to 500 in increments of 10, and from 500 to 900 in increments of 100. Similarly, the number of estimators in gradient boosted models vary from one to 200, in increments of 10. C-index values for a range of estimators for RSF, gradient boosted model with regression tree base leaners and with least squares base learners using SMOTE balancing are given in Figure 4.6, 4.7, and 4.8, respectively.

a) b)

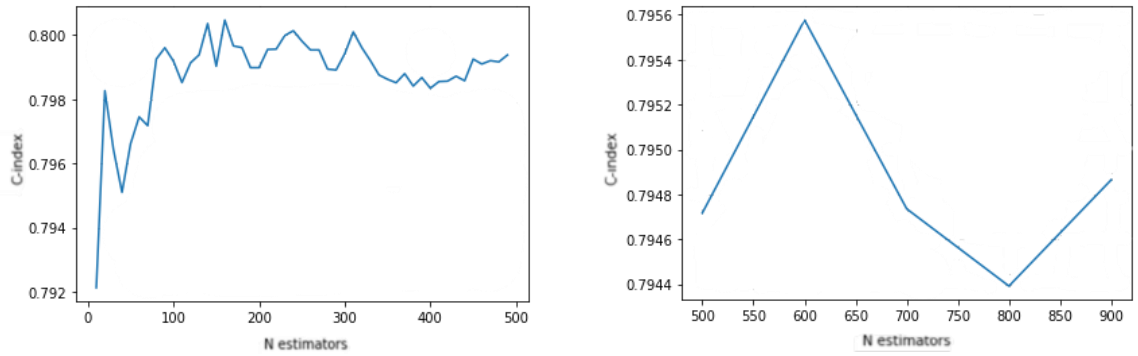


Figure 4.6: C-index values corresponding to the number of estimators ranging (a) between [10,500], and (b) [500,900] for a random survival forest model

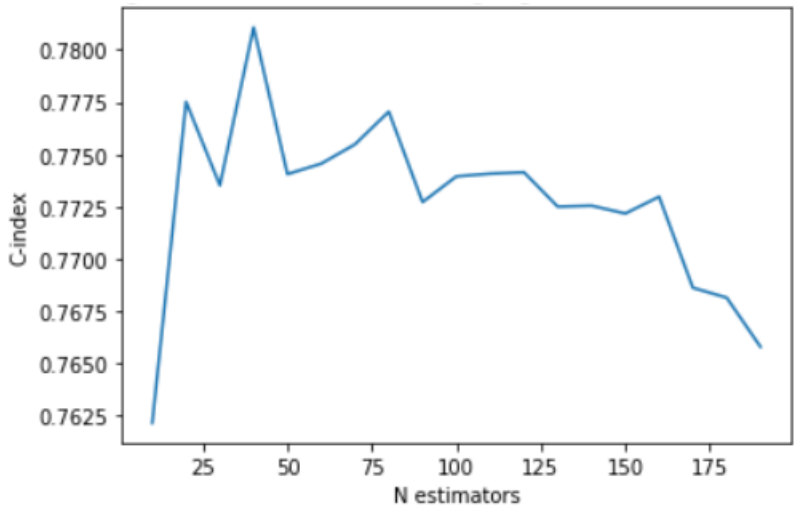


Figure 4.7: C-index values corresponding to the number of estimators in a regression tree base learner gradient boosted model

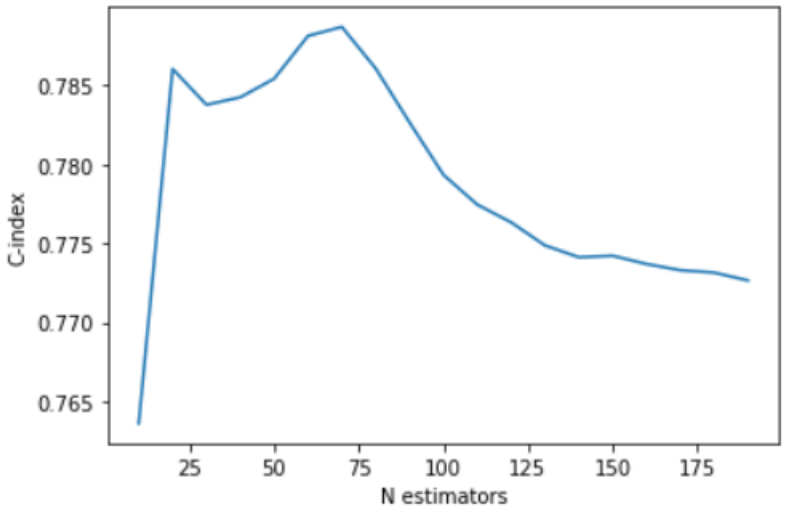


Figure 4.8: C-index values corresponding to the number of estimators in a least square base learner gradient boosted model

4.2.2 Concordance Index Results

Briefly, the C-index is a goodness of fit statistic that is used to measure the degree of agreement between the risk score generated and the time-to-failure. This is a metric that evaluates the predicted output of the model; a C-index value of one indicates perfect predictive accuracy of the test data. Table 4.2 summarizes the C-index results from the permutations of class balancing techniques and survival models. The C-index values in Table 4.2 are generated from the models with optimized hyperparameters as described in Figures 4.6-4.8 above and for models that do not require hyperparameter optimization.

Results for Unbalanced Data

The C-index of survival models trained on the unbalanced dataset indicate that the models perform strongly, given the data presented [25, 68]. RSF outperforms all models, with a C-index value of 0.8321. This includes outperforming the two-stage Cox PHM/RSF model slightly ($\mathrm{C-index} = 0.8311$), indicating that the reduction of covariates does not improve the C-index of the model. This is because RSF was designed for scaling up with high-dimensional data and a large sample size [67].

The Cox PHM model generates a C-index value of 0.5- equivalent to that of flipping a coin. Several factors may contribute to the poor performance of the model. Computation of Cox PHM requires the inverse of a matrix, which only exists if the matrix is of full rank. With the increased number of covariates and instances, columns may be linearly dependent on one another, resulting in a matrix that is potentially not full rank. The model then requires a penalization parameter, α, for regularization to penalize values in attempt to maintain a full rank matrix and generate a C-index value and avoid model errors. Aside from the high dimensional nature of the data, at some level the covariates used are likely to be correlated, which contributes to co-linearity issues commonly found in Cox PHM [141]. Issues pertaining to the model could have also arisen if the proportional hazards assumption is violated; this assumes that all instances have the same relative hazard function but differ only in the function's scaling factor. These are but several potential explanations for the poor performance of Cox PHM.

The success of gradient boosted model lies in the base learner used. Given that RSF is fundamentally an ensemble regression tree, using regression tree base learners produces C-index values larger than those of Cox PHM and Survival SVMs. Regression tree base learner gradient boosting outperforms the least squares base learner gradient boosted model, with respect to C-index for unbalanced data. The regression tree base learner model uses target values rather than threshold value constraints which simplifies the problem using equality constraints whereas least squares base learners operate by minimizing the sum of squared errors. By minimizing the squared error, the model effectively reduces misclassification rates to the best of the model's ability.

Kernel survival SVMs indicate slightly lower C-index values when compared with other models, with the sigmoid kernel providing the poorest accuracy among the entire test set, for both unbalanced and unbalanced data[1]. The linear survival SVM outperforms all kernel survival SVMs, generating a slightly higher C-index than both the third-degree polynomial and the sigmoid kernel survival SVMs. This result concurs with the notion that the dataset is more linearly separable when described in the feature space compared to other kernel-based survival

[1] This excludes the Cox PHM results ($\mathrm{C-index} = 0.500$) due to the inherent flaw within the model's ability to bypass assumptions and correctly interpret the input data.

SVMs. A kernel density plot is derived from the test data used for survival SVMs, shown in Figure 4.9. The kernel density plot visualizes the dispersion of test data over cable age. Peaks in the distribution at certain cable ages indicate where the number of instances is most concentrated. The dispersion of failed and active cables, from Figure 4.9, indicate that the instances can be largely separated with a linear kernel with minimal overlap, except at a cable age of 45 years.

The third-degree polynomial and sigmoid kernel survival SVMs are non-parametric, and their complexity grows with the size of the training dataset. Computationally speaking, it is more expensive to compute the kernel survival SVMs and requires the projection of the data into a higher dimensional space where the data can then be linearly separable. Tuning the hyperparameters in the kernel survival SVMs is also extremely tedious, with more hyperparameters requiring tuning. The resultant is a model that is overfit, with lesser accuracy.

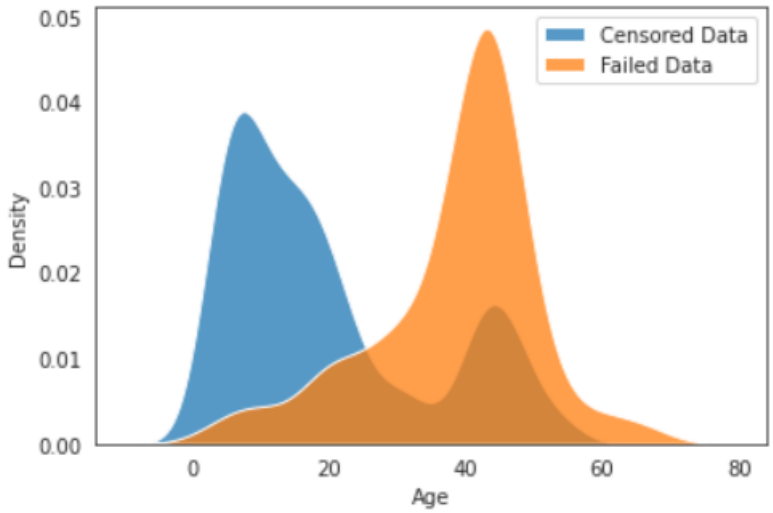


Figure 4.9: Kernel density plot for survival SVM test data

Results for Balanced Data

The C-index results for various class balancing methods indicate several notable features seen in Table 4.2. Firstly, based on the C-index, RSF outperforms or performs on par with all other models in all class balancing techniques, with only slightly lesser values than the two-stage model with Tomek Links ($\text{RSF C} - \text{index} = 0.8311$) and random over-sampling ($\text{RSF C} - \text{index} = 0.8083$)- a difference in C-index of 0.0003 and 0.0039 for the two models,

respectively. The findings that RSF performs greater than or equal to the other survival models with respect to C-index concurs with the results from other studies [1, 11, 67].

Second, Cox PHM using random under-sampling, random over-sampling, and SMOTE do not possess the same issues associated with that of the unbalanced, Tomek Links, and NearMiss models in which the C-index results are equivalent to random guessing. C-index values for Cox PHM with random under-sampling, random over-sampling, and SMOTE are 0.768, 0.765, and 0.769, respectively.

The results of gradient boosted models and survival SVMs, both linear and kernel, follow the same patterns as described for the models using the unbalanced dataset indicating consistency in the testing methods and further, the consistency in the results.

Table 4.2: Test data concordance index results

Model Method/ Imbalance Technique	**RSF**	**Cox PHM**	**Two-Step Approach**	**Gradient Boosted Cox w/ Regression Tree Base Learner**	**Gradient Boosted Cox w/ Least Square Base Learner**	**Linear Survival SVM**	**Linear Kernel Survival SVM**	**3rd Degree Polynomial Kernel Survival SVM**	**Sigmoid Kernel Survival SVM**
Original	0.8321	0.5000*	0.8311	0.8095	0.7958	0.7699	0.7241	0.7239	0.5014
RUS	0.8092	0.7679	0.8013	0.7839	0.7867	0.7482	0.7240	0.7239	0.5008
Tomek Links	0.8311	0.5000*	0.8314	0.8075	0.7964	0.7875	0.7241	0.7239	0.5008
NearMiss	0.8079	0.5000*	0.7906	0.7565	0.7948	0.7636	0.7240	0.7239	0.5000
ROS	0.8083	0.7653	0.8122	0.7800	0.7857	0.7513	0.7240	0.7239	0.5519
SMOTE	0.8096	0.7694	0.8017	0.7696	0.7841	0.7467	0.7240	0.7239	0.7198

The results from Table 4.2 seemingly indicate that the models that perform most accurately in terms of C-index values are those in which the models operate on unbalanced data, with the RSF model outperforming all methods with all balancing techniques. Naive selection, however, of a model based on a single performance metric does not provide enough information to determine the best model for the dataset; hence, the analysis of the integrated Brier score.

4.2.3 Integrated Brier Score Results

While the results of the C-index provide discriminative power, i.e., ranking C-index scores to compare models, the analysis of integrated Brier score (IBS) provides insight into both calibration and discrimination power of the model. Recall, an IBS value $\in [0,1]$ closer to zero indicates perfect calibration and discrimination ability and the perfect overall performance of the survival models.

The C-index results indicate that the accuracy of models used on unbalanced data is greatest, the IBS values prove otherwise. The resultant IBS values found for each model are outlined in Table 4.3. It should be prefaced that survival SVMs can only predict a relative risk score and not a probability, therefore the IBS cannot be computed for survival SVMs and is thus omitted in the analysis [139].

The unbalanced RSF and two-stage models produce C-index values of 0.8321 and 0.8311, however the IBS values are 0.0392 and 0.4000, respectively; greater than the IBS values found using the same models that are balanced, implying a better fitting model for the dataset when employing class balancing in the pre-processing stage of the method in terms of classification and discrimination power, i.e., the IBS.

Models that produced the best IBS, i.e., the lowest scores, include SMOTE with gradient boosted regression tree base learners and SMOTE-RSF, IBS $= 0.0263$ and 0.0261, respectively, as well as random over-sampling with a regression tree gradient boosted model (IBS $= 0.0260$) and random under-sampling with Cox PHM (IBS $= 0.0265$).

Issues arising with the C-index in certain Cox PHM balancing methods, i.e., Tomek Links and NearMiss, as well as Cox PHM with unbalanced data also translate into greater IBS values indicative less predictive performance and calibration of these models.

Other models incorporating Tomek Links class balancing produce IBS values similar to those of models with unbalanced data. NearMiss also produces high IBS scores compared to other class balancing methods. The former is likely due to the nature of Tomek Links and the

inability to provide a truly balanced dataset for the model's learning stage. Hence, the results are similar to that of the unbalanced dataset, where the proportion of censored samples remains far greater than that of the uncensored samples. NearMiss operates similar to Tomek Links in that majority class instances are removed to create greater dispersion between opposite classes. This could be indicative of the IBS results generated by NearMiss models being similar to Tomek Links balancing models and unbalanced data models. Both of the class balancing methods experience the same phenomenon as the unbalanced dataset in the Cox PHM.

Table 4.3: Test data integrated Brier score results

Model Method/ Imbalance Technique	RSF	Cox PHM	Two-Step Approach	Gradient Boosted Cox w/ Regression Tree Base Learner	Gradient Boosted Cox w/ Least Square Base Learner	Linear Survival SVM	Linear Kernel Survival SVM	3rd Degree Polynomial Kernel Survival SVM	Sigmoid Kernel Survival SVM
Original	0.0392	0.0517	0.0400	0.0398	0.0412	-	-	-	-
RUS	0.0274	0.0265	0.0275	0.0299	0.0272	-	-	-	-
Tomek Links	0.0379	0.0487	0.0382	0.0378	0.0390	-	-	-	-
NearMiss	0.0335	0.0412	0.0363	0.0315	0.0296	-	-	-	-
ROS	0.0289	0.0267	0.0312	0.0260	0.0269	-	-	-	-
SMOTE	0.0261	0.0279	0.0270	0.0263	0.0283	-	-	-	-

4.2.4 Processing Speed

To gauge model performance given the dataset and the ability to scale up where additional data may be implemented, the run time, averaged across five runs, is computed under the SMOTE balancing techniques. The SMOTE technique is used as it provides the largest training dataset, with the largest number of unique instances; recall, techniques such as random over-sampling duplicate data instances rather than creating synthetic points as in SMOTE. The run times, determined on a 4 core, 8 GB RAM, 4.0 GHz processing speed workstation, are presented in Table 4.4.

The Cox PHM and linear survival SVM models have the quickest run time by a large margin, compared to all other survival models. The most accurate models in terms of C-index and IBS, i.e., RSF and the gradient boosted models, have a longer run time, with RSF requiring

approximately twice as much time to formulate a model compared with the gradient boosted model with regression tree base learners.

Table 4.4: Survival model average run time

Survival Model	Run Time
RSF	6 min 35 sec
Cox PHM	11 sec
2-Step Approach	10 min 21 sec
Gradient Boosted Model w/ Reg. Tree	3 min 23 sec
Gradient Boosted Model w/ LS	12 min 15 sec
Linear Survival SVM	6 sec
Kernel Survival SVM	2 min 32 sec

4.2.5 Variable Importance

Variable importance is determined for the survival models given the respective covariates used in the dataset. The results in Table 4.5 are determined using the RSF model with SMOTE class balancing. Positive scores indicate that the C-index score decreases when the feature is removed and therefore the feature contributes to the accuracy of the model. Scoring weights close to zero contain little-to-no useful information in the model and may be omitted from the model with minimal consequence. The weight associated with covariates refers to the average decrease of the test data C-index, and conversely the average increase of the C-index if the removed feature weight is positive and negative, respectively. For example, removing the cable length covariate from the datasets, on average, decreases the C-index by 0.0989 points

From Table 4.5, the cable length is of most importance to the construction of the model, followed by cable diameter, number of splices, and number of cables in the configuration. Features that contain little-to-no importance to the model include cables that are in a duct line arrangement, as well as various insulating materials. This is due to the minimal number of instances of these properties when compared to their counterparts, recall Figure 4.2 – 4.4.

Table 4.5: Feature importance and weighting of covariates

Weight	Feature
0.0989 ± 0.0229	GEOMETRIC_LENGTH
0.0555 ± 0.0139	DIAM
0.0334 ± 0.0100	NUM_SPLICES
0.0154 ± 0.0146	NUM_CABLES
0.0153 ± 0.0070	INSULATION_XLPE CN
0.0095 ± 0.0027	INSULATION_TRXLPE
0.0088 ± 0.0041	MATERIAL_CU
0.0066 ± 0.0075	MATERIAL_AL
0.0019 ± 0.0021	ARRANGEMENT_DB
0.0004 ± 0.0007	ARRANGEMENT_Ductline
0.0000 ± 0.0002	INSULATION_PILC
0.0000 ± 0.0000	INSULATION_EPR LC Shield-TRIPLEX
0 ± 0.0000	INSULATION_XLPE LC Shield-TRIPLEX
-0.0000 ± 0.0001	INSULATION_EPR CN

To further the understanding of variable importance, the SMOTE with Cox PHM model is analyzed for the *log hazard ratio* that corresponds to the β value of the covariate. The log hazard ratios of the covariates are summarized in Table 4.6. Positive log hazard ratio suggests increased risk associated with the covariate on the cable's failure likelihood, whereas negative log hazard ratios indicate a smaller risk.

Notable comparisons can be drawn when comparing various cable properties. Cables with aluminum conducting material (LHR = 0.5314) possess a greater risk factor when compared with copper conducting material (LHR = -0.6760). While aluminum conducting material is lighter in weight and more susceptible to bending for installation, copper conducting material has a lower impedance and can carry a higher capacity, making copper conductors slightly smaller than aluminum conductors of the same current capacity.

Cables that are directly buried (LHR = 0.1451) are more prone to failure than cables that are encased in a duct line (LHR = -0.0969). This is because directly buried cable is more likely to experience thermal bottlenecking due to soil characteristics and thermal resistivity. Organics in the soil impact the ability of the cables heat removal. Duct lines, on the other hand, are a more consistent environment, where the surrounding medium is air and, in some cases, water. Duct

lines also provide the ability for inspection through manholes which also presents the early identification of issues arising in the cable.

Increasing the number of cables in the configuration, i.e., one-phase, two-phase, and three-phase, has a positive effect on the survival of the distribution cable in the network (LHR = -0.2896).

Insulation material comprising of cross-linked polyethylene with a concentric neutral (XLPE CN), are of smaller risk (LHR = -0.2975) than PILC cables (LHR = 0.2463). The remaining insulation types are infrequently present in the dataset and thus, lack of information pertaining to these insulating materials is the contributing factor to their respective log hazard ratio.

Table 4.6: Covariate log hazard ratios

Feature	Log Hazard Ratio
INSULATION_TRXLPE	3.803
MATERIAL_AL	0.5314
INSULATION_PILC	0.2463
DIAM	0.1684
ARRANGEMENT_DB	0.1451
NUM_SPLICES	0.1252
GEOMETRIC_LENGTH	0.0006
ARRANGEMENT_DUCTLINE	-0.0969
NUM_CABLES	-0.2896
INSULATION_XLPE CN	-0.2975
MATERIAL_CU	-0.6760
INSULATION_EPR CN	-10.58
INSULATION_EPR LC Shield-TRIPLEX	-10.61
INSULATION_XLPE LC Shield-TRIPLEX	-10.77

4.3 Underlying Distribution

The empirical CDF generated from the SMOTE-RSF model is used as the basis of the K-S test. Formulating the empirical CDF, $F(t)$, is done by utilizing the final iteration of the survival function, $S(t)$,

$$F(t) = 1 - S(t) \tag{4.1}$$

where the survival function is determined through the hazard function, $H(t)$,

$$S(t) = \exp(-\mathrm{H(t)}) \tag{4.2}$$

The resultant CDF is graphically represented in Figure 4.10. The empirical CDF gives way to firstly identifying notable theoretical distribution through the overall shape and form of the function, visually. Then the data used to create the empirical CDF forms the basis of the K-S test, where the data is mapped against theoretical distributions constructed from random continuous variable states and the shape and form of the function is numerically compared.

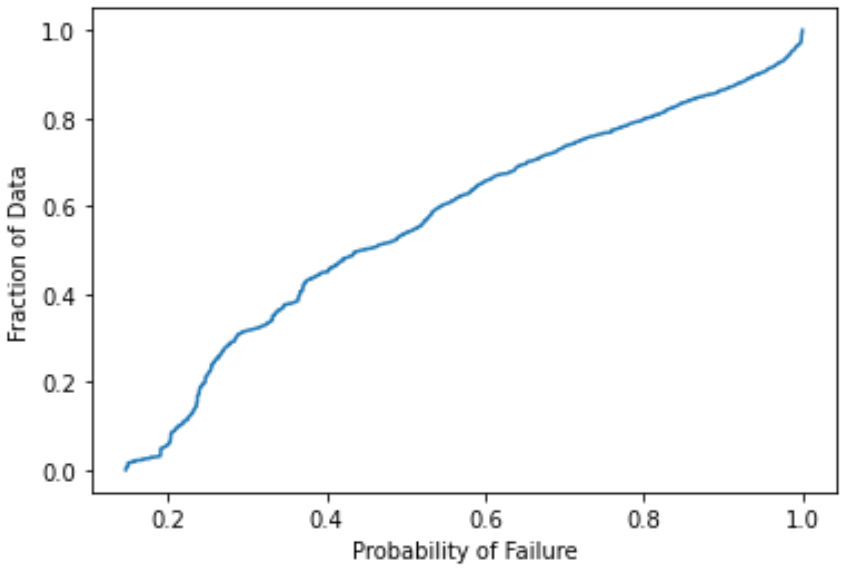


Figure 4.10: Empirical CDF generated by RSF cumulative survival function

The empirical CDF is evaluated with several theoretical CDFs including the Weibull, exponential, normal, and chi squared distributions, outlined in Table 4.7. Using a 95% confidence interval, i.e., $\alpha = 0.05$, and the number of samples in the training dataset, $N = 9776$, the critical value, d, is calculated as follows:

$$d \approx \frac{1.36}{\sqrt{N}} \tag{4.3}$$

$$d \approx \frac{1.36}{\sqrt{9776}} = 0.014$$

From the p-values in Table 4.7, it can be said that the several tests reject the null hypothesis, $p < 0.05$, that the samples are drawn from the same distribution. This includes the Weibull test, the exponential test where $shape \neq 0$ and $scale \neq 1$, and the Chi squared test where the degree of freedom ≥ 4. The K-S test critical value derived above (Equation 4.3) is used to evaluate the remaining K-S tests. Critical values generated from the test, D, that are of significance are *less than* the calculated critical values, $d = 0.014$. The hypothesis that the empirical CDF is drawn from the sample distribution is *rejected* if $D > 0.014$. The critical values of the remaining theoretical distributions that are not initially rejected based on p-value are all rejected for their critical value output.

The results from the K-S test indicate that the empirical CDF generated from the survival model does not conform to any population distribution. Thus, it is concluded that the empirical distribution of the study is not generated from a specific underlying distribution.

Table 4.7: K-S test critical values and p-value results

Theoretical Distribution	Parameters	Critical Value	p Value
Weibull	Shape = 1	0.992	1.24e-06
Exponential	Shape = 0	0.667	0.074
	Scale = 1		
	Shape ≠ 0	1.00	6.10e-10
	Scale ≠ 1		
Normal	-	0.562	0.205
Chi Squared	DoF = 1	0.461	0.432
	DoF = 2	0.667	0.074
	DoF = 3	0.737	0.037
	DoF = 4	0.952	0.0002

4.4 Individualized Hazard and Survival Functions

While ensemble hazard and survival functions give a broad understanding of the average cable's failure likelihood, individualized hazard and survival functions of cable instances enable a more expansive understanding of specific cables, with specific covariate conditions. Generation of the individualized asset functions, first, requires the model to be able generate ensemble hazard and survival functions, which requires estimation of the probability of the ensemble failure likelihood. For this reason, survival SVMs are unable to produce individualized asset functions.

Of the combination of class balancing methods and survival models, the SMOTE-RSF model was extended from the analysis in Section 4.3 based on the high C-index and low IBS value to be used to create individualized survival and hazard functions. Other models, such as Cox PHM and gradient boosted models are also capable of individualized hazard and survival function generation using alternative technique of estimation.

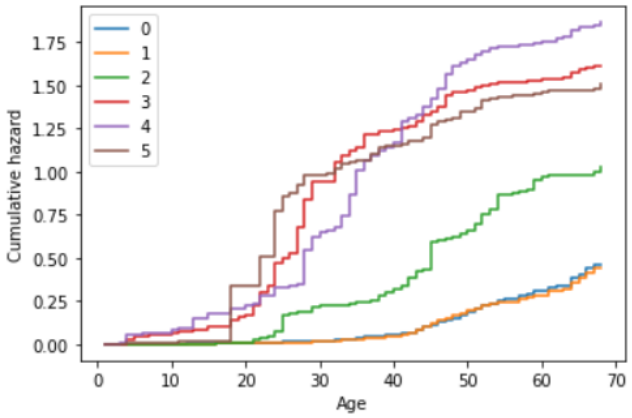
The model is fit to the test data which are arranged in sequential order based on the total length of the cables. From this, the three shortest and three longest cables are selected for the prediction of the individualized survival and hazard functions. The data selected for graphical representation is shown in Table 4.8.

Table 4.8: Cable test data arranged by shortest cable to longest cable

Legend	Diameter	Length (m)	Number of Cables	Number of Splices	Conducting Material	Insulating Material	Burial Arrangement
0	8.25	0.219	3	0	Al	XLPE CN	Directly Buried
1	8.25	0.598	2	0	Al	XLPE CN	Directly Buried
2	17.96	0.615	3	0	Cu	XLPE CN	Directly Buried
3	17.96	1058.508	3	0	Al	XLPE CN	Directly Buried
4	15.03	1170.131	3	1	Al	XLPE CN	Directly Buried
5	22.00	1578.222	3	0	Cu	XLPE CN	Directly Buried

The SMOTE-RSF method produces hazard and survival functions, depicted in Figure 4.10. Results from the estimation indicate that cables of shorter length (Instances 0-2) have a higher survival probability and lower hazard with an increase in age. This concurs with the findings in Verweij et al. [134] in that longer cables possess a greater hazard than shorter cables.

a)



b)

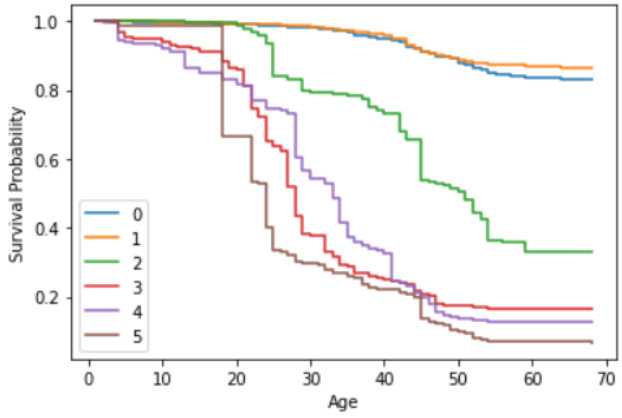


Figure 4.11: (a) Hazard functions and (b) Survival functions of cable instances based arranged by cable length generated from SMOTE-RSF

4.5 Summary of Results

The results of the study are analyzed and discussed in a sequential fashion as per the workflow created for the assessment and comparative analysis of survival modelling of underground distribution cables.

Class balancing techniques are implemented, and their balancing methods and end results are discussed. This leads into the formulation of the various survival models used with each class balancing technique to generate performance measures that can be compared to determine the models that are best suited for the use on the given dataset. The underlying distribution analysis is computed and no relation to population distributions is found that accepts the hypothesis that the empirical distribution is drawn from a population distribution.

The hypotheses laid out in the study are validated at various stages of the modelling procedure. The relative importance of variables on the outcome of survivability and their respective impact on the accuracy of the model is determined through feature weights and log hazard ratios to satisfy the first hypothesis. The second hypothesis is validated through the use of the integrated Brier score, in which unbalanced modelling possesses less calibration and discrimination power when compared with balanced models. To satisfy the third hypothesis, the C-index and the integrated Brier score are used to compare the performance of the models. This hypothesis is violated with the use of survival SVMs, which do not possess the ability to generate an integrated Brier score due to the nature of the model. The fourth and final hypothesis is validated using the SMOTE-RSF model. The model is selected based on the high C-index result and low IBS value, i.e., a model that is among the best fit the underground cable dataset. The SMOTE-RSF model produces stepwise hazard and survival functions corresponding to three of the shortest and three of the longest cables in the test dataset.

Chapter 5

Conclusions

This chapter reviews the steps that were followed in testing the hypotheses for a method of conducting a comparative survival analysis for medium-voltage underground distribution cable.

There was a comprehensive examination of cable survival data, from which came class balancing techniques and the development of survival models for a data-driven survival analysis. Multiple methods were implemented for creating predictive models for underground cables. These methods were reviewed for their discrimination and calibration performance and were extended to survival and hazard functions for individual cable instances.

5.1 Development of Comparative Analysis Methods

The comparative survival analysis method required a sequential approach to modelling. Firstly, covariates that were appropriate for providing a detailed description of the cable data were required. Covariates were selected by identifying cable characteristics and operating conditions that were likely to increase the failure likelihood of a cable based on expert knowledge as well as the availability of the data pertaining to the covariates. Since the effect on model performance at an individualized level was of concern, multiple cable characteristics and environmental conditions were assessed in the context of survival function prediction of underground cables.

Due to the highly imbalanced nature of real-world survival data, class balancing techniques were included in the pre-processing stage of the modelling procedure. Both under-sampling and over-sampling methods were explored for their impact on the accuracy of the survival models when compared with the results from the unbalanced dataset. Desired features of a model were: a high accuracy (C-index) and low integrated Brier score; the ability to determine

feature importance of covariates on the model; and the extension of ensemble survival and hazard functions to cables on an individualized basis. With that, survival models were selected on the basis of the study's objective. Well known, simple survival models for survival analysis were explored and models that are more complex, formed as extensions to simpler models, were also tested. Employing a variety of modelling methods enabled a greater understanding of how increasing the complexity of a model effected the calibration and discrimination performance of the model given a high-dimensional dataset.

5.1.1 Review of Methods

The methods employed in the study consist of a variety of class balancing and survival model alternatives, each of which presented advantages and drawbacks in their own respect.

Including High-Dimensional Data in the Model

The high-dimensional nature of the data presented an added level of complexity to the modelling methods employed. While models that possess more data are able train with more information, often resulting in higher accuracy, the larger the number of covariates, the larger the chance of the covariates being dependent. While still preserving the amount of information contained in the models, careful consideration was given to the selection of covariates in the study. Covariates that contain full data, i.e., no missing entries, covariates that were seemingly independent of one another, and covariates that encompass as much explanatory information about the cable and its operating environment as possible were selected.

Choice of Class Balancing

Class balancing came in the form of under-sampling methods, namely, random under-sampling, Tomek Links and NearMiss, and over-sampling methods including random over-sampling and SMOTE. The selected methods were chosen based on their widespread use in balancing data as well as their proven capability to effectively deal with large datasets.

Alternatives to class balancing techniques were also considered, however, they were not pursued. The alternatives included: changing performance metrics resampling the dataset; and

employing cost-sensitive training. Changing the performance metric would not allow for both calibration and discrimination to be evaluated and recall survival analysis is a subset of regression where the common metrics of regression do not apply to the objectives of the study. Resampling the data would effectively be manipulating the data to achieve a higher performance, however, the objective of the study was to allow the models to build themselves without any external aid and determine how they perform. Finally, a cost-sensitive model was not employed because of the practical implementation challenges associated with the size of the data and the overall function of this modelling method is more so for classification rather than regression and survival analysis.

5.1.2 Comparing Survival Models

Limitations of Survival Analysis with Unbalanced Data

Although survival analysis can be conducted on the original dataset with no set pre-processing requirement, highly unbalanced data poses a challenge that can seemingly not be overcome without additional measures. Models trained on unbalanced data, where the number of censored instances far outweighs the number of uncensored instances, often undervalue the information associated with the minority (uncensored) class. These instances are viewed as "noise" in the modelling system and are often predicted incorrectly. This is seen in the calibration power of the model, in the form of the integrated Brier score.

Selection of Survival Models

The main objective of this study was determining the best suited survival model for the underground cable data. Both semi-parametric and non-parametric survival models were studied to examine the influence of assumptions of the shape of the functional relationship between dependent variables and covariates versus models that are free of any assumptions and allow the model to develop a unique relationship between event variables and the covariates used.

The survival models were selected on the basis of their previous use in survival analysis studies; their ability to transform complex datasets into usable and useful information related to

the objectives of the study; and their proven ability to be highly accurate models for large and high-dimensional datasets.

It was therefore determined that the following models would be implemented in the study:

- Random survival forests
- Cox proportional hazards model
- Two step Cox and random survival forest model
- Gradient boosted Cox model with regression tree base learners
- Gradient boosted Cox model with least squares base learners
- Linear survival support vector machine
- Linear kernel survival support vector machine
- Polynomial kernel survival support vector machine
- Sigmoid kernel survival support vector machine

Using Performance Metrics as a Comparative Tool

The relative performance of the models and their respective class balancing methods across the various model and class balancing techniques was gauged via performance metrics. The performance metrics encompassed both the model's discrimination power and calibration. In doing so, the models were comparable on a consistent basis, with performance metrics that could be directly compared against one another.

The concordance index was the primary performance metric, which was a reflection of the model's accuracy in measuring the agreement between the risk score and the time-to-event. The concordance index, more specifically, Harrell's concordance index, was used due to its widespread applicability to all survival analyses. The limitation of the use of concordance index was that first, models cannot be evaluated on the basis of a single metric for comparison purposes, and second, the model's calibration level was not evident when using this metric. Thus, the integrated Brier score was used as a secondary metric that would allow for the understanding models discrimination and calibration power. The caveat of the integrated Brier score is that it is calculated based on the survival function; this became problematic in survival

analysis when regression and classification-based models that are extended to survival analysis were unable to generate hazard and survival functions, i.e., survival support vector machines.

5.2 Simulation

The simulation conducted consisted of a four step, methodological approach, to ultimately predict the survival and hazard functions of individual cable instances. The first step included data collection and data pre-processing. Cable data was collected for 8172 cable instances, which included cables that were actively in-service as well as cable that experienced failure. Among these cables, data pertaining to the physical cable properties and operating environment were extracted for use as covariates, or explanatory variables. A total of nine covariates, inclusive of the dependent variables, i.e., event class status and time elapsed (age) was included in the study. Among these covariates, three were deemed categorical in nature and required additional processing measures to convert string-based information into usable, value-based, variables. This was processed using one-hot encoding in which covariates were split into a binary metric based on the presence or lack thereof, of that covariate category; the resultant was a 8172 x 16 dataset.

From here, the data was split into training data and testing data, where class balancing techniques were applied to the training data in order to balance the data. The unbalanced training dataset was made up of 4888 active cables and 587 failed cables. Under-sampling and over-sampling methods were both used to balance the number of censored (active) and uncensored (failed) cable instances; in the former, the output yielded a training data set with 587 active and 587 failed cables, with the exception of Tomek Links, which did not truly balance the dataset, rather, removed instances of active cable that were nearest-neighbours of failed cables. The results of the Tomek Links yielded a training dataset that consisted of 4747 active cables and 587 failed cables. In the latter sampling method, the training dataset was expanded to 4888 active and 4888 failed cable instances either using random duplication or creating synthetic samples of the minority class.

With the data being successfully pre-processed, the second step involved employing the reformatted and rebalanced data into various survival models. Each survival model was

individually evaluated using the unbalanced dataset and all methods of class balancing. The models that required hyperparameter tuning were optimized based on the concordance index values by adjusting the number of estimators of the model. Model execution was done with a one stage approach, except for the two stage Cox and random survival forests model. In this model, the first step involved running the Cox proportional hazard model and obtaining the feature importance values of the covariates in the study. From this, five covariates were selected based on the log hazard ratio values and those covariates were used for a random survival forest simulation as opposed to the complete covariate set used in the one stage models.

The models were optimized via the concordance index, which led to the third step of the method. The optimized concordance index was recorded and additionally, the integrated Brier score was computed for the models. It was noted that the Brier score, and hence, the integrated Brier score, was formed based on the computed survival function generated from the model; models that did not generate a survival function were unable to output an integrated Brier score, namely survival support vector machines.

The final step in the method was generating individualized survival curves for cable instances; this was done using SMOTE-RSF. The output determined that individual survival instances were in fact able to be created from the ensemble survival function, with stepwise survival and hazard functions generated for six of the test data cables.

5.3 Assessment of Hypotheses

This section assesses whether the hypotheses for the comparative survival analysis of underground cables have been validated.

5.3.1 Hypothesis One

The first hypothesis to be validated was that the properties of a cable and the operating environment have an observable effect on failure of a cable. This hypothesis was satisfied by both feature importance and through the log hazard ratio. The former was developed to better

understand the covariates that held the most weight in the outcome of the model's output. The latter determined the increase, decrease, or lack of effect of the covariates and their influence on the hazard of the cables.

5.3.2 Hypothesis Two

The second hypothesis was that there are positive effects on the performance of survival models with the inclusion of class balancing techniques. The counter hypothesis was there was no performance increase in the models with the presence of balanced data. The counter hypothesis was refuted by analyzing the integrated Brier score, which determined that the models' discrimination and calibration levels increased with the presence of balanced data.

5.3.3 Hypothesis Three

The third hypothesis to validate was that there existed performance metrics that allowed for the comparison of models on the same basis. The concordance index was calculated for all models with and without class balancing techniques. The concordance index was calculated from fitting the model to unseen test data and did not pose any restrictions on the methodology as all models were fit to the respective test data.

The integrated Brier score, however, was not calculated for survival support vector machines. The output of the survival support vector machines was in the form of a relative risk score of cable instances, whereas the requirement of determining the integrated Brier score is derived from the survival function, generated from probabilities, and not risk scores. Thus, hypothesis three was partially disproven with the use of survival support vector machines.

5.3.4 Hypothesis Four

The fourth, and final, hypothesis was that the ensemble survival models could be related to cables on an individual instance level. SMOTE-RSF was used based on its high concordance index and low integrated Brier score to validate the hypothesis. The SMOTE-RSF model generated stepwise survival and hazard functions pertaining to individual cables. Due to their inability of generating probabilities, ensemble survival and hazard functions, and therefore

individualized survival and hazard functions, survival support vector machines do not validate the fourth hypothesis.

5.4 Contributions of the Present Work

This work presents original contributions to the application of survival analysis in a comparative fashion for real-world data. Methods for comparing survival models with class balancing techniques were critically evaluated. For modelling purposes, it was appropriate to assess the viability of class balancing in survival analysis, and so the inclusion was compared with unbalanced data modelling.

A methodology was developed for identifying the efficacy of class balancing with multiple survival models, and a process for comparing the accuracy and goodness-of-fit of the models was developed. The methods created enable the evaluation of covariate importance and their influence of the model output.

The present work thus contributes to both comparative assessments of time-to-event models and reliability analytics.

Chapter 6

Recommendations

Recommendations for future work include expanding the breadth of information of physical and environmental parameters for more accurate risk scoring, additional simulations to ensure that the results described above are not limited to the models used, and validation of results using current in-situ cable testing procedures.

The goal of future efforts is a highly accurate reliability analysis applicable to real-world applications.

6.1 Data Collection

6.1.1 Expanding Covariate Information

Since the key to reliability analytics and predictive maintenance is to gather information about the system of concern, there exists a need for more avenues of data collection from data generated, and more means of processing.

It would not be onerous to include additional cable covariates into the modelling procedure. Current datasets in use can offer additional information that include more expansive geotechnical information, cable installation information, and time-dependent cable operating conditions; information pertaining to load conditions can even be included as a time-dependent covariate [34].

An alternative communication medium for the extraction of covariate information is a direct link from Microsoft Azure and Databricks to the model. In doing so, a more concise and real-time dataset can be implemented directly into the model for a more seamless integration between the modelling methods and the data collected. Python has the capability to connect

directly with both Azure and Databricks [32, 93], making the process more applicable for this integration. The merits of different database communication methods should be investigated.

A data issue that should be examined is the storage of distribution cable information. Covariate information was obtained from various sources and databases; issues arose in the duplication of data and where different datasets contained the same data information. The creation of a unified and all-inclusive database for the storage of underground cable information would greatly improve the efficiency to construct a dataset for reliability modelling. Maintenance and upkeep of information then becomes an artifact within the data pre-processing ability rather than separate from it. The requirements of such a database, and its continuous monitoring, remain future work.

6.1.2 Implementation of Time-Dependent Covariates

Operating conditions can be monitored in near real time, with a continuous feed of data being generated for specific characteristics. Soil moisture content, water table fluctuation, temperature, and other related conditions are monitored through external sources, with multiple monitoring locations around the distribution cable network. These are considered time-dependent covariates [76]. Consumer power use, monitored with granularity between 15 minutes and 60 minutes, would also provide valuable insight into the electrical throughput of the distribution network downstream. The state of the system, i.e., whether cables are energized or de-energized (negligible current), would be an important metric to further explore; this would help determine whether the constant re-energization cycles impact the reliability of cables [94]. Theoretically, data is available to estimate the real time current for every cable segment; however, with the lack of data processing and formation of a collective data pool as mentioned, the use of real time current calculations is not currently possible.

The addition of time-dependent covariates allows for greater information to be processed in the model in conjunction with the time-independent covariates examined in this study [146]. In doing so, continually updated probabilities for cable failure can be achieved to simulate the variations experienced by active cable, to better answer the questions of why a specific cable may have failed given that the time-independent covariates indicated that there was a low level

of failure probability and how intermittent overloading and underloading conditions effect a cable's failure likelihood.

While they do exist, the number of models capable of dealing with time-dependent covariates are far fewer than those which operate on time-independent covariates [142]. Different performance metrics would likely be required to capture the dependency of the model on time. The type of modelling required to encompass such data would create additional levels of complexity to the formulation and are thus out of scope, remaining as future work.

6.2 Real World Validation

Simulating the effects of cable characteristics and the operating environment on the likelihood of failure is one step to enhancing the reliability and understanding the hazards on the cables in a distribution network. With an understanding of the risk scores and failure probabilities, the focus can be shifted to cables with the highest failure probability. In doing so, annual cable testing procedures can focus on the high-risk cables, and the performance of these cables can be monitored to verify the results from the developed model. Testing procedures for the high-risk cables should encompass a time-domain reflectometry [73] test for detection of potential fault locations in cable and splice segments, a tan delta test [106] diagnosing insulation quality and degradation, and partial discharge test [124] for identifying electrical discharges in cable segments that contribute to increased cable deterioration. These tests detect a wide range of deterioration and failure mechanisms and provide insight into the health and quality of underground distribution cables.

Including the results from the survival models on an individual asset level into the annual testing plan with the appropriate testing procedures will validate, or invalidate, the results concluded. This remains future work for the utility provider.

6.3 A Step Toward Improved Cable Reliability

Improved cable reliability through predictive maintenance demands supplements for cable testing procedures. The procedures offered in this work allows for supplementation to cable reliability understanding in a real-world application, a step toward that goal.

References

[1] Adham, D., Abbasgholizadeh, N., and Abazari, M., "Prognostic factors for survival in patients with gastric cancer using a random survival forest.," *Asian Pacific journal of cancer prevention: APJCP,* vol. 18, no. 1, p. 129, 2017.

[2] Alghamdi, A. S. and Desuqi, R. K., "A study of expected lifetime of XLPE insulation cables working at elevated temperatures by applying accelerated thermal ageing," *Heliyon: e03120,* vol. 6, no. 1, 2020.

[3] Altman, D. G. and Bland, J. M., "Time to event (survival) data," *Bmj,* vol. 317, no. 7156, pp. 468-469, 1998.

[4] Ashrit, L., "Underground Power Cables – Architecture, Applications and Advantages," ed: Electricalfundablog.

[5] Barandela, R., M, Valdovinos R., S., Sánchez J., and J., Ferri F., "The imbalanced training sample problem: Under or over sampling?," in *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*. Berlin, Heidelberg: Springer, 2004, pp. 806-814.

[6] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C., "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explorations,* vol. 6, no. 1, pp. 20-29, 2004.

[7] Bendell, A., "Proportional hazards modelling in reliability assessment," *Reliability Engineering,* vol. 11, no. 3, pp. 175-183, 1985.

[8] Bennin, K. E., Phannachitta, P., Monden, A., and Mensah, S., "Mahakil: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction," *IEEE Transactions on Software Engineering,* vol. 44, no. 6, pp. 534-550, 2017.

[9] Bhagat, R. C. and Patil, S. S. (2015) Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data using Random Forest. *2015 IEEE International Advance Computing Conference (IACC)*. 403-408.

[10] Bloom, J. A., Feinstein, C., and Morris, P. (2006) Optimal replacement of underground distribution cables. *2006 IEEE PES power systems conference and exposition*. 389-393.

[11] Bohannan, Z. S., Coffman, F., and Mitrofanova, A., "Random survival forest model identifies novel biomarkers of event-free survival in high-risk pediatric acute lymphoblastic leukemia," *Computational and structural biotechnology journal,* vol. 20, pp. 583-597, 2022.

[12] Botchkarev, A., "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," *arXiv preprint arXiv,* vol. 1809, no. 03006, 2018.

[13] Breiman, L., "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[14] Brier, G. W., "Verification of forecasts expressed in terms of probability," *Monthly weather review,* vol. 78, no. 1, pp. 1-3, 1950.

[15] Brownlee, J., "Why one-hot encode data in machine learning," *Machine Learning Mastery,* pp. 1-46, 2017.

[16] Buhari, M., Levi, V., and Awadallah, S. K., "Modelling of ageing distribution cable for replacement planning," *IEEE Transactions on Power Systems,* vol. 31, no. 5, pp. 3996-4004, 2015.

[17] Buhlmann, P. and Yu, B., "Boosting with the L 2 loss: regression and classification," *Journal of American Statistical Association,* vol. 98, no. 462, pp. 324-339, 2003.
[18] Carlin, C. S. and Solid, C. A., "An apporach to addressing selection bias in survival analysis," *Statistics in Medicine,* vol. 33, no. 23, pp. 4073-4086, 2014.
[19] Chang, H. H., Reich, B. J., and Miranda, M. L., "Time-to-event analysis of fine particle air pollution and preterm birth: results from North Carolina, 2001–2005," *American journal of epidemiology,* vol. 175, no. 2, pp. 91-98, 2012.
[20] Chawla, N. V. and al., et, "SMOTE: synthetic minority over-sampling technique.," *Journal of artificial intelligence research,* vol. 16, pp. pp321-357, 2002.
[21] Chen, Y., Jia, Z., Mercola, D., and Xie, X., "A gradient boosting algorithm for survival analysis via direct optimization of concordance index," *Computational and mathematical methods in medicine,* 2013.
[22] Clark, T., Bradburn, M., Love, S, and , et al., "Survival Analysis Part I: Basic Concepts and First Analyses," *British Jounral of Cancer,* vol. 89, pp. 232-238, 2003.
[23] Collett, D., *Modelling survival data in medical research*. CRC press, 2015.
[24] Cortes, C. and Vapnik, V., "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.
[25] Cox, D. R. and Oakes, D., *Analysis of Survival Data*. Chapman and Hall/CRC, 2018.
[26] Cox, D. R., "Regression Models and Life Tables," *Journal of the Royal Statistical Society,* vol. 34, no. 2, pp. pp.187-220, 1972.
[27] Csanyi, E., "Copper or aluminum? Which one to use and when?," ed: Electrical Engineering Portal (EEP), 2012.
[28] D'Agostino, R. B. and Stephens, M. A., "Goodness-of-fit Techniques," New York: Marcel Dekker, 1986.
[29] D., Li., Xuan, G., Shan, Z., and al., et, "Application of Cox PHM in coagulation function analysis in pateints with primary liver cancer," *Journal of Xi'an Jiaotong University (Medical Sciences),* vol. 31, no. 1, pp. 99-103, 2012.
[30] Dag, A., Oztekin, A., Yucel, A., Bulur, S., and Megahed, F. M., "Predicting heart transplantation outcoes through data analytics," *Decision Support Systems,* vol. 94, pp. 42-52, 2017.
[31] Dal Pozzolo, A., Caelen, O., Johnson, R. A., and Bontempi, G., "Calibrating probability with undersampling for unbalanced classification," in *2015 IEEE Symposium Series on Computational Intelligence*: IEEE, 2015, pp. 159-166.
[32] Databricks, "Connect Python and pyodbc to Databricks," ed, 2022.
[33] Datta, G., E., Alexander L., A., Hinterberg M., and Hagar, Y., "Balanced Event Prediction Through Sampled Survival Analysis," *Systems Medicine,* vol. 2, no. 1, pp. 28-38, 2019.
[34] Dekker, F. W., De Mutsert, R., Van Dijk, P. C., Zoccali, C., and Jager, K. J., "Survival analysis: time-dependent effects and time-varying risk factors," *Kidney International,* vol. 74, no. 8, pp. 994-997, 2008.
[35] Dietrich, S., Floegel, A., Troll, M., and al., et, "Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis.," *International journal of epidemiology,* vol. 45, no. 5, pp. 1406-1420, 2016.
[36] Domingues, I., Amorim, J. P., Abreu, P. H., and al., et, "Evaluation of Oversamplig Data Balancing Techniques in the Context of Ordinal Classification," *2018 International Journal Conference on Neural Networks,* vol. IEEE, pp. 1-8, 2018.

[37] Dong, X., Yuan, Y., Gao, Z., and al., et. (2014, June) nalysis of cable failure modes and cable joint failure detection via sheath circulating current. *2014 IEEE Electrical Insulation Conference (EIC)*. 294-298.

[38] Doshi, D. A., Khedkar, K. B., Raut, N. T., and Kharde, S. R., "Real time fault failure detection in power distribution line using power line communication," *International Journal of Engineering Science,* vol. 4834, 2016.

[39] Dreiseitl, S. and Ohno-Machado, L., "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics,* vol. 35, no. 5-6, pp. 352-359, 2002.

[40] Elhassan, T. and Aljurf, M., "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method," *Global J Technol Optim S,* vol. 1, 2016.

[41] Elith, J., Leathwick, J. R., and Hastie, T., "A working guide to boosted regression trees," *Journal of Animal Ecology,* vol. 77, no. 4, pp. 802-813, 2008.

[42] Evers, L. and Messow, C. M., "Sparse kernel methods for high-dimensional survival data," *Bioinformatics,* vol. 24, no. 14, pp. 1632-1638, 2008.

[43] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D., "Regression models," in *Regression*. Berlin, Heidelberg: Sringer, 2021, pp. 23-84.

[44] Fantazzini, D. and Figini, S., "Random survival forests models for SME credit risk measurement.," *Methodology and computing in applied probability,* vol. 11, no. 1, pp. 29-45, 2009.

[45] Farhadian, M., Dehdar Karsidani, S., Mozayanimonfared, A., and Mahjub, H., "Risk factors associated with major adverse cardiac and cerebrovascular events following percutaneous coronary intervention: a 10-year follow-up comparing random survival forest and Cox proportional-hazards model," *BMC Cardiovascular Disorders,* vol. 21, no. 1, pp. 1-8, 2021.

[46] Fouodo, C. J., König, I. R, , Weihs, and al., et, "Support Vector Machines for Survival Analysis with R," *R Journal,* vol. 10, no. 1, 2018.

[47] Friedman, J. H., "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of statistics,* pp. 1189-1232, 2001.

[48] Friedman, J. H., Hastie, T., and Tibshirani, R., "Additive Logistic Regression: A statistical View of Boosting (with discussion and a rejoinder by the authors)," *The annals of statistics,* vol. 28, no. 2, pp. 337-407, 2000.

[49] Frisk, E., Krysander, M., and Larsson, E., "Data-driven lead-acid battery prognostics using random survival forests.," *In Annual Conference of the PHM Society,* vol. 6, no. 1, 2014.

[50] Gel, Y., Raftery, A. E., and Gneiting, T., "Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method," *Journal of the American Statistical Association,* vol. 99, no. 467, pp. 575-583, 2004.

[51] Gémar, G., Moniche, L., and Morales, A. J., "Survival analysis of the Spanish hotel industry," *Tourism Management,* vol. 54, pp. 428-438, 2016.

[52] George, B., Seals, S., and Aban, I., "Survival analysis and regression models," *Journal of nuclear cardiology,* vol. 21, no. 4, pp. 686-694, 2014.

[53] Goli, S., Mahjub, H., Faradmal, J., Mashayekhi, H., and Soltanian, A. R., "Survival prediction and feature selection in patients with breast cancer using support vector regression," *Computational and mathematical methods in medicine,* 2016.

[54] González-Domínguez, J., Sánchez-Barroso, G., García-Sanz-Calcedo, J., and de Sousa Neves, N., "Cox Proportional Hazards Model Used for Predictive Analysis of the Energy Consumption of Healthcare Buildings," *Energy and Buildings,* vol. 257, 2022.
[55] Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M., "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine,* vol. 18, no. 17, pp. 2529-2545, 1999.
[56] Guo, S., *Survival Analysis*. Oxford University Press, 2010.
[57] Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A., "Regression modeling strategies for improved prognostic prediction," *Statistics in Medicine,* vol. 3, no. 2, pp. 143-152, 1984.
[58] Harrell, F. E., Califf, R. M., and al., et, "Evaluating the yield of medical tests," *Jama,* vol. 247, no. 18, pp. 2543-2546, 1982.
[59] Harrell Jr, F. E., Lee, K. L., and Mark, D. B., "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.," *Statistics in medicine,* vol. 15, no. 4, pp. 361-287, 1996.
[60] Hasib, K. M. and al., et, "A survey of methods for managing the classification and solution of data imbalance problem," *Journal of Computer Science,* vol. 16, no. 11, pp. p1546-1557, 2020.
[61] He, K., Li, Y., Zhu, J., and al., et, "Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates," *Bioinformatics,* vol. 32, no. 1, pp. 50-57, 2016.
[62] Hougaard, P., "Fundamentals of Survival Data," *International Biometric Society,* vol. 55, no. 1, pp. 13-22, 1999.
[63] Hsu, C. W, Chang, C. C., and J., Lin C., "A practical guide to support vector classification," pp. 1396-1400, 2003.
[64] Hummel, T. J. and Sligo, J. R., "Empirical comparison of univariate and multivariate analysis of variance procedures," *Psychological bulletin,* vol. 76, no. 1, p. 49, 1971.
[65] Ishaq, A., Sadiq, S., Umer, M., and al., et, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access,* vol. 9, pp. 39707-39716, 2021.
[66] Ishwaran, H., Kogalur, U. B.., Blackstone, E. H., and Lauer, M. S., "Random Survival Forests," *The Annals of Applied Statistics,* vol. 2, no. 3, pp. 841-860, 2008.
[67] Ishwaran, H. and Kogalur, U. B., "Random survival forests for R," *R News,* vol. 7, no. 2, pp. 25-31, 2007.
[68] Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J., "Random Survival Forests for High Dimensional Data," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 4, no. 1, pp. 115-132, 2011.
[69] Jamet, B., Morvan, L., Nanni, C., and al., et, "Random survival forest to predict transplant-eligible newly diagnosed multiple myeloma outcome including FDG-PET radiomics: a combined analysis of two independent prospective European trials," *European journal of nuclear medicine and molecular imaging,* vol. 48, no. 4, pp. 1005-1015, 2021.
[70] Jeatrakul, P., Wong, K. W., and Fung, C. C., "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm.," in *Springer*, Berlin, Heidelberg, 2010, pp. 152-159.

[71] Jenkins, S. P., "Survival Analysis," *Unpublished manuscript, Institute for Social and Economic Research,* vol. 42, pp. 54-56, 2005.
[72] Jo, T. and Japkowicz, N., "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter,* vol. 6, no. 1, p. 40, 2004.
[73] Jones, S. B., Wraith, J. M., and Or, D., "Time domain reflectometry measurement principles and applications," *Hydrological processes,* vol. 16, no. 1, pp. 141-153, 2002.
[74] Justel, A., Peña, D., and Zamar, R., "A multivariate Kolmogorov-Smirnov test of goodness of fit," *Statistics & probability letters,* vol. 35, no. 3, pp. 251-259, 1997.
[75] Kalbfleisch, J and Prentice, R., *The Statistical Analysis of Failure Time Data*. New York: Wiley, 1980.
[76] Kartsonaki, C., "Survival analysis," *Diagnostic Histopathology,* vol. 22, no. 7, pp. 263-270, 2016.
[77] Klerx, M., Morren, J., and Slootweg, H., "Analyzing Parameters that Affects the Reliability of Low Voltage Cable Grids and their Applicability in Asset Management," *IEEE Transactions on Power Delivery,* vol. 34, no. 4, pp. 1432-1441, 2019.
[78] Kotsiantis, S., Kanellopoulos, D., and Pintelas, P., "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering,* vol. 30, no. 1, pp. 25-36, 2006.
[79] Kvamme, H. and Borgan, Ø., "The brier score under administrative censoring: Problems and solutions," *arXiv preprint arXiv:1912.08581,* 2019.
[80] Lane, W. R., Looney, S. W., and Wansley, J. W., "An application of the Cox proportional hazards model to bank failure," *Journal of Banking & Finance,* vol. 10, no. 4, pp. 511-531, 1986.
[81] Learn, Imbalanced, "Under-sampling methods," ed: Imbalanced learn.
[82] Lemaitre, G., Nogueira, F., and Aridas, C. K., "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research,* vol. 18, no. 17, pp. 1-5, 2017.
[83] Li, H. and Luan, Y., "Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data," *Bioinformatics,* vol. 21, pp. 2403-2409, 2005.
[84] Liu, X. Y. and Zhou, Z. H., "The influence of class imbalance on cost sensitive learning: An Empirical Study," in *Sixth International Conference on Data Mining (ICDM'06)*, Washington, D.C., USA, 2006: IEEE, pp. 970-974.
[85] Longato, E., Vettoretti, M., and Di Camillo, B., "A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models," *Journal of Biomedical Informatics,* vol. 108, 2020.
[86] Lopes, R. H., Reid, I. D., and Hobson, P. R., "The two-dimensional Kolmogrov-Smirnov test," 2007.
[87] Machin, D. and Cheung, Y. B., Parmar, M., *Survival Analysis: A Practical Approach*. John Wiley & Sons, 2006.
[88] Mani, I. and Zhang, I., "kNN approach to unbalanced data distributions: a case study involving information extraction," *Proceddings of workshop on learning from imbalanced datasets. ICML,* vol. 126, pp. 1-7, 2003.
[89] Massey Jr., F. J., "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American statistical Association,* vol. 46, no. 253, pp. 68-78, 1951.

[90] McCarthy, K., Zabar, B., and Weiss, G. M., "Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes?," *Proceedings of the 1st international workshop on Utility-based data mining,* pp. 69-77, 2005.
[91] McDonald, B., "Aluminum Vs. Copper Conductors." [Online]. Available: https://www.helukabel.com/publication/us/trade-mag-articles/wind-systems-july-2017-aluminum-vs-copper.pdf
[92] Miao, F., Cai, Y. P., Zhang, Y. X., and al., et, "Risk prediction of one-year mortality in patients with cardiac arrhythmias using random survival forest.," *Computational and mathematical methods in medicine 2015,* 2015.
[93] Microsoft, "Python on Azure," ed.
[94] Montanari, G. C., Seri, P., Bononi, S. F., and Albertini, M., "Partial discharge behavior and accelerated aging upon repetitive DC cable energization and voltage supply polarity inversion," *IEEE Transactions on Power Delivery,* vol. 36, no. 2, pp. 578-586, 2020.
[95] Montanari, G. C., Cavallini, A., and Puletti, F., "A new approach to partial discharge testing of HV cable systems," *IEEE electrical insulation magazine,* vol. 22, no. 1, pp. 14-23, 2006.
[96] Newson, R. B., "Comparing the predictive powers of survival models using Harrell's C or Somers' D," *The Stata Journal,* vol. 10, no. 3, pp. 339-358, 2010.
[97] Ng, W. W., Hu, J., Yeung, D. S., Yin, S., and Roli, F., "Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems," *IEEE transactions on cybernetics,* vol. 45, no. 11, pp. 2402-2412, 2014.
[98] Nguyen, N. P., *Gradient Boosting for Survival Analysis with Applications in Oncology*. University of South Florida, 2019.
[99] Ohno-Machado, L., "Modeling medical prognosis: survival analysis techniques," *Journal of biomedical informatics,* vol. 34, no. 6, pp. 428-439, 2001.
[100] Oladunni, T., Tossou, S., Haile, Y., and Kidane, A., "COVID-19 County Level Security Classification with Imbalanced Class: A NearMiss Under-sampling Approach," *medRxiv,* 2021.
[101] Panmala, N., Suwanasri, T., and Suwanasri, C. (2020, June) Condition assessment of medium voltage underground cable systems. *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. 210-213.
[102] Park, S. Y., E., Park J., Kim, H., and Park, S. H., "Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches)," *Korean Journal of Radiology,* vol. 22, no. 10, p. 1697, 2021.
[103] Pecorelli, F., Di Nucci, D., De Roover, C., and De Lucia, A., "On the Role of Data Balancing for Machine Learning-Based Code Smell Detection," in *Proceedings of the 3rd ACM SIGSOFT international workshop on machine learning techniques for software quality evaluation*, 2019.
[104] Pölsterl, S., Navab, N., and Katouzian, A. (2015, September) Fast training of support vector machines for survival analysis. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 243-259.
[105] Pölsterl, S., Navab, N., and Katouzian, A., "An efficient training algorithm for kernel survival support vector machines," *arXiv preprint arXiv:1611.07054,* 2016.

[106] Ponniran, A. and Kamarudin, M. S., "Study on the performance of underground XLPE cables in service based on tan delta and capacitance measurements," in *2008 IEEE 2nd International Power and Energy Conference*, 2008.
[107] Provost, F., "Machine Learning from Imbalanced Data Sets 101," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2008.
[108] Rackwitz, R., "Reliability analysis—a review and some perspectives," *Structural safety,* vol. 23, no. 4, pp. 365-395, 2001.
[109] Razali, N. M. and Wah, Y. B., "Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests," *Journal of statistical modeling and analytics,* vol. 2, no. 1, pp. 21-33, 2011.
[110] Retterath, B., Venkata, S. S., and Chowdhury, A. A. (2004, September) Impact of time-varying failure rates on distribution reliability. *2004 International Conference on Probabilistic Methods Applied to Power Systems*. 953-958.
[111] Ridgeway, G., "The state of boosting," *Computing science and statistics,* pp. 172-181, 1999.
[112] Rodríguez-Torres, F., Martínez-Trinidad, J. F., and Carrasco-Ochoa, J. A., "An Oversampling Method for Class Imbalance Problems on Large Datasets," *Applied Sciences,* vol. 12, no. 7, p. 3424, 2022.
[113] Rosner, B., Glynn, R. J., and Ting Lee, M. L., "Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach," *Biometrics,* vol. 59, no. 4, pp. 1089-1098, 2003.
[114] Satpathy, S., "Overcoming Class Imbalance using SMOTE Techniques," ed: Analytics Vidhya, 2020.
[115] Schapire, R.E., "The strength of weak learnabilit," *Machine Learning,* vol. 5, pp. 197-227, 1990.
[116] Schmid, M., Wright, M. N., and Ziegler, A., "On the use of Harrell's C for clinical risk prediction via random survival forests," *Expert Systems with Applications,* vol. 63, pp. 450-459, 2016.
[117] Schröer, G. and Trenkler, D., "Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples," *Computational statistics & data analysis,* vol. 20, no. 2, pp. 185-202, 1995.
[118] Shivaswamy, P. K., Chu, W., and Jansche, M. (2007, October) A support vector approach to censored targets. *Seventh IEEE international conference on data mining (ICDM 2007)*. 655-660.
[119] Steennis, E. F. and Kreuger, F. H., "Water treeing in polyethylene cables," *IEEE Transactions on Electrical Insulation,* vol. 25, no. 5, pp. 989-1028, 1990.
[120] Sujatha, R., Chatterjee, J. M., Jhanjhi, N. Z., and al., et, "Heart Failure Patient Survival Analysis with Multi Kernel Support Vector Machine," *INTELLIGENT AUTOMATION AND SOFT COMPUTING,* vol. 32, no. 1, pp. 115-129, 2022.
[121] Sun, X., Lee, W. K., Hou, Y., and Pong, P. W. T., "Underground Power Cable Detection and Inspection Technology Based on Magnetic Field Sensing at Ground Surface Level," *IEEE Transactions on Magnetics,* vol. 50, no. 7, pp. 1-5, 2014.
[122] Sun, Y., Wong, A., and Kamel, M. S., "Classification of Imbalanced Data: A Review," *International Journal of Pattern Recognition and Artifical Intelligence,* vol. 23, no. 4, pp. 687-719, 2009.

[123] Tang, Z., Zhou, C., Jiang, W., and al., et, "Analysis of Significant Factors on Cable Failure Using Cox Proportional Hazards Model," *IEEE Transactions on Power Delivery,* vol. 29, no. 2, pp. 951-957, 2013.
[124] Tian, Y., Lewin, P. L., and Davies, A. E., "Comparison of on-line partial discharge detection methods for HV cable joints," *IEEE Transactions on Dielectrics and Electrical Insulation,* vol. 9, no. 4, pp. 604-615, 2002.
[125] Tomek, I., "Two Modifications of CNN," *IEEE Transactions on Systems Man and Communications,* vol. 6, pp. p769-772, 1976.
[126] Trepanier, M., Tremblay, C., Reynaud, L., and Lachance, M., "New Life for Underground." [Online]. Available: https://www.omicronenergy.com/en/news/details/ensuring-the-reliability-of-underground-cable-systems/#
[127] Tyagi, S. and Mittal, S., "Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning," *Lecture Notes in Electrical Engineering,* vol. 597, 2020.
[128] Uno, H., Cai, T., Pencina, M. J., and al., et, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in medicine,* vol. 30, no. 10, pp. 1105-1117, 2011.
[129] Valle, Y.D., Hampton, N., Perkel, J., and Riley, C., "Underground Cable Systems," in *Meyers, R.A. (eds) Encyclopedia of Sustainability Science and Technology*. New York, NY.: Springer, 2012.
[130] Van Belle, V., Pelckmans, K., Van Huffel, S., and Suykens, J. A. K., "Support vector methods for survival analysis: a comparison between ranking and regression approaches," *Artificial Intelligence in Medicine,* vol. 53, no. 2, pp. 107-118, 2011.
[131] Van Belle, V., Pelckmans, K., Suykens, J. A., and Van Huffel, S. (2007, July) Support vector machines for survival analysis. *Proceedings of the third international conference on computational intelligence in medicine and healthcare (cimed2007)*. 1-8.
[132] Van Belle, V., Pelckmans, K., Suykens, J.A.K., and Van Huffel, S., "Survival SVM: a practical scalable algorithm," *16th European Symposium on Artificial Neural Networks,* pp. 89-94, 2008.
[133] Vapnik, V., "The support vector method of function estimation," *Nonlinear modeling,* pp. 55-85, 1998.
[134] Verweij, R., van Houwelingen, D., and Prein, A., "Where To Replace Assets? Spatial Analysis on Differntial Ageing of Low-Voltage PILC Cables," *CIRED-Open Access Proceedings Journal,* vol. 1, pp. 2585-2589, 2017.
[135] Weiss, G. M., "Mining with rarity: a unifying framework," *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 7-19, 2004.
[136] Widodo, A. and Yang, B. S., "Machine health prognostics using survival probability and support vector machine," *Expert Systems with Applications,* vol. 38, no. 7, pp. 8430-8437, 2011.
[137] Wilkins, D. J., "The Bathtub Curve and Product Failure Behavior," ed: Weibull, Reliability HotWire, 2002.
[138] Witten, D. M. and Tibshirani, R., "Survival Analysis with High-Dimensional Covariates," *Statistical methods in medical research,* vol. 19, no. 1, pp. 29-51, 2010.
[139] Xiao, J., Mo, M., Wang, Z., and al., et, "The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study," *JMIR medical informatics,* vol. 10, no. 2: e33440, 2022.

[140] Xu, H. and Sinha, S. K., "Modeling pipe break data using survival analysis with machine learning imputation methods," *J. Perform. Constr. Facil,* vol. 35, no. 5, 2021.
[141] Xue, X., Kim, M. Y., and Shore, R. E., "Cox regression analysis in presence of collinearity: an application to assessment of health risks associated with occupational radiation exposure," *Lifetime data analysis,* vol. 13, no. 3, pp. 333-350, 2007.
[142] Yao, W., Frydman, H., Larocque, D., and Simonoff, J. S., "Ensemble methods for survival function estimation with time-varying covariates," *Statistical Methods in Medical Research,* p. 09622802221111549, 2022.
[143] Yu, D. J., Hu, J., Tang, Z. M., Shen, H. B., Yang, J., and Yang, J., "Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling," *Neurocomputing,* vol. 104, pp. 180-190, 2013.
[144] Zdilar, L., Vock, D. M., and Marai, et al., "Evaluating the effect of right-censored end point transformation for radiomic feature selection of data from patients with oropharyngeal cancer," *JCO clinical cancer informatics,* vol. 2, pp. 1-19, 2018.
[145] Zhang, Y. and Haghani, A., "A Gradient Boosting Method to Improve Travel Time Predictions," *Transportation Research Part C: Emerging Technologies,* vol. 58, pp. 308-324, 2015.
[146] Zucker, D. M. and Karr, A. F., "Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach," *The Annals of Statistics,* vol. 18, no. 1, pp. 329-353, 1990.
[147] Zuech, R., Hancock, J., and Khoshgoftaar, T. M., "Detecting web attacks using random undersampling and ensemble learners," *Journal of Big Data,* vol. 8, no. 1, pp. 1-20, 2021.