ARCADA

Scheduling of preventive maintenance using prognostic models

A case study on elevator doors

Mikko Alutoin

Master's Thesis

Master of Engineering - Big Data Analytics

2020

MASTER'S THESIS	
Arcada	
Degree Programme:	Big Data Analytics
Identification number:	
Author:	Mikko Alutoin
Title:	Scheduling of preventive maintenance using prognostic models - A case study on elevator doors
Supervisor (Arcada):	Dr. Leonardo Espinosa
Commissioned by:	Dr. Olli Mali (KONE Oyj)

Abstract:
Aim of this thesis is to research how maintenance process of elevator doors can be optimised. To this end, a business goal is set. It is to decrease the amount of unplanned maintenance visits caused by door malfunctions. A corresponding analytics goal is defined, which is used for ranking elevators that would make best candidates for maintenance in this respect. Thesis shows that scheduling of preventive maintenance visits using prognostic models would be beneficial in reducing the rate of unplanned maintenance visits.
The result was obtained via a practical case study using a real-life dataset containing data for more than 20 thousand elevators for a period of two years. Major part of the work was to form this dataset from (partially incomplete) raw data that consisted of various maintenance records and condition monitoring data on elevator doors. Tested models were the well-known Cox Proportional Hazards model, and a more recent recurrent neural network model called Weibull Time to Event RNN (WTTE-RNN).
Survival analysis methods were used to extract information from partial observations. Prior to training the models, stratified survival curves were obtained via Kaplan-Meier estimator for two groups: all elevators and freshly maintained elevators. Difference in these curves quantifies the difference that preventive maintenance visit generally yields. Then, prognostic models were used for producing daily predictions of the survival curve for each elevator. Elevators were ranked using these predictions and based on how much their condition seemed to have worsened over time (as this is thought to capture their potential to benefit from maintenance). Finally, a comparison between highly ranked elevators over all elevators is provided to demonstrate the skill of the models.

Keywords:	Kone, data science, neural network, prognosis, maintenance companies, sliding doors
Number of pages:	53
Language:	English
Date of acceptance:	25.2.2020

CONTENTS

Figures

Tables

FOREWORD

This thesis was commissioned by my advisor and superior Dr. Olli Mali (KONE Oyj). I would like to express my gratitude to him for the idea, trust, good spirit, and companionship. Without his continuous support the thesis would never have been started yet alone completed. Also, I am thankful for the resources admitted to me by Kone Corporation, and for my colleagues who are the most driven group of professionals and fun to work with.

Special thanks goes to my supervisor and lecturer Dr. Leonardo Espinosa (Arcada), who kept me going even when it became evident that the initial approach to the problem wasn't working. His encouragement and true curiosity towards my research was astonishing. I owe him (and Olli) for the proof reading and many suggestions that made this thesis a better read.

I would also like to thank Dr. Magnus Westerlund (Arcada), the programme director of the master's degree programme in Big Data Analytics. My studies in Arcada have opened a new world to me professionally. Finally, I am also grateful to my lecturer Dr. Anton Akusok, who steered me in the right direction in the beginning of this thesis work, when I struggled to find a good approach and methods for the research problem.

Writing this thesis has taught me a lot and I hope you'll find the topic worthwhile the effort.

Hyvinkää, February 2020

Mikko Alutoin

1 INTRODUCTION

The topic of this thesis and addressed research problems are introduced in this chapter. It begins by discussing a setting where a company is offering preventive maintenance and corrective maintenance as a continuous service to its customers. Possibilities of machine learning for this maintenance business process automation and optimisation are described and a business goal is set. Then more detailed analytics goals are considered which helps to formulate the research questions and limit the scope of the thesis. The introduction is concluded by an outline of the structure of the thesis.

1.1 Business goal

Artificial intelligence and machine learning techniques are in the rise across nearly all industries. Although many machine learning techniques, such as neural networks, have existed for a long time, it seems that only recently have they gained enough momentum to make a significant impact to our daily lives. The most prominent evidence of this are the totally new products that are being introduced by some of the top technology companies. For instance, self-driving vehicles are an example of this category where technological limits are being pushed further by machine learning. Nevertheless, these types of applications are only the tip of an iceberg. Majority of data scientists today are applying machine learning in a more conventional setting where development can be described more as an evolution than revolution: existing digital services are being optimised and automated through machine learning and advanced analytics. Companies are jumping on the wagon and rethinking the ways of conducting their business processes and services via harvesting and analysing data and leveraging insights from data analytics.

This thesis focuses on the latter type of application, more specifically automation of maintenance business process of an industrial service provider. The service provider is offering a continuous maintenance service for elevators. Condition monitoring data is continuously collected from elevators that are under maintenance contract. There exists also data about all maintenance visits. The aim of the thesis is to research how the collected data can be used to optimise the maintenance process to achieve a business goal. By maintenance process we refer to the process in which maintenance technicians are

dispatched to conduct site visits for performing preventive and corrective maintenance actions. There are different types of visits. Majority of the visits are preventive maintenance visits. These are scheduled and recurring visits where maintenance technician does preventive checks for the machinery, and performs certain maintenance actions, such as lubrication of elevator parts. In many countries the frequency of preventive visits is governed by the law, in order to guarantee safe operation of the elevator. Another type of visit is a repair visit when certain elevator parts are replaced. Trigger for replacement may have originated due to the part having reached its predetermined end of life, technician noticing the need during preventive maintenance visit, or customer having reported a problem with the equipment. Finally, there are so-called unplanned visits. These visits are needed when an unprecedented problem has occurred with the elevator. Either the end customer or facility manager has contacted the help desk, or fault has been detected automatically (e.g. by means of data analytics). In this case a maintenance technician shall visit the site immediately to resolve the technical problem. Technician needs to cease any actions he may be conducting at the time and travel to the site where his help is urgently needed. Entrapment of a passenger inside the elevator car is an example of most urgent type of unplanned visit, which must be responded in a timely manner despite the hour.

Optimising any business process begins with defining clear business goals. Only after the business goals and more specific analytics goals have been clarified it is time to apply suitable methods, such as machine learning. Let us now proceed by defining the business goal. It is evident that the cost associated with an unplanned site visit is very high. Direct cost follows from the unprecedented nature of the site visit that causes inefficiency to the otherwise streamlined process. Indirect cost is associated with the perceived quality of the maintenance service by the customer. Customer satisfaction is in jeopardy whenever an unplanned visit is called for, since often that is accompanied by decreased serving time of the elevator. We arrive to our business goal definition:

Business Goal: Decrease the rate of unplanned maintenance visits that are caused by door malfunctions.

This thesis focuses on door malfunctions as a lot of condition monitoring data is available on elevator doors and door malfunctions represent a large portion of unplanned visits. There are basically two ways to address the business goal using data driven analytics (given that the rate of preventive maintenance visits is not to be increased). On one hand the preventive maintenance visits may be scheduled more optimally to reduce the rate of failures. Idea being, that preventive maintenance actions themselves may keep the machinery in good enough condition or that maintenance technician is able to detect more timely that a repair visit is called for. The other path to improvement is to provide maintenance technicians with insights of the root cause to the failure, so that they would be able to remedy the problem more efficiently (e.g. avoid multiple visits to the site for correcting the fault). This thesis focuses on the former path by studying how scheduling of preventive maintenance visits could be based on recommendations from prognostic models.

1.2 Analytics goal

Authors in Nalchigar (2018a) lay out a framework where an analytics project is described, conducted, and documented via three complementary views called 1) business view, 2) analytics design view, and 3) data preparation view. This framework provides a useful separation of concerns and is therefore used throughout this thesis also. In business view, stakeholders document their requirements for the business process as a whole and set requirements for the analytics design view. The business requirements lead to setting one or more analytics goals. Role of the analytics goal is to provide insights from data to be used by various actors. The data preparation view will be discussed later in Chapter 3. Let us now discuss the analytics design view and define the analytics goal (given the business goal defined in Section 1.1).

Capabilities of different machine learning algorithms may be used in fulfilling the business goal. This means providing insights through some analytics goal, such as Prediction Goal, Description Goal, or Prescription Goal (Nalchigar 2018a). For instance, unsupervised machine learning may be used to provide anomaly detection (Description Goal). In the context of this thesis, analysis could be done on different elevator rides and outliers could be detected which in turn could facilitate finding and solving prob-

lems effectively. Alternatively, supervised machine learning algorithms could be applied to predict the emergence of such anomalies or any undesirable departure from targeted service level in advance (Prediction Goal). For example, a supervised machine learning could be trained for evaluating risk of a malfunction using maintenance records as the ground truth. Finally, a supervised machine learning model could even provide maintenance technician with a list of most probable causes for a potential or already reported malfunction (Prescription Goal). This would require that root cause analysis for past malfunctions would be available as ground truth in training.

Analytics goal constitutes a bridge between business goal and machine learning model. It is the assignment for the modelling exercise in the analytics design view. What model is applicable is determined by available data. Without ground truths, for instance, it is not possible to use supervised machine learning. The quantity of the data plays a role too - more data supports more complex algorithm. In case the analytics goal is a prediction goal, it essentially specifies the target variable (i.e. what is to be predicted). Analytics goal should be simple and allow definition of concise and measurable research questions. In this thesis the primary analytics goal is:

Analytics Goal: Predict the value of a preventive maintenance visit (with respect to the business goal).

Note that this is a pure prediction goal. Specifically, it is not in the scope of this thesis to provide a prescriptive model for helping the maintenance technicians with the root cause analysis. Instead the aim is to build a supervised machine learning model that can be used to answer the following question: “If a preventive maintenance visit was to be scheduled on a particular elevator today, what would be the expected benefit with respect to the business goal of reducing the rate of door malfunctions?”

Rationale behind the selected analytics goal is that there must exist an optimal sequence for performing the preventive maintenance visits. This sequence is the one that minimises the number of unplanned visits caused by door malfunctions. Predicting probability of a door malfunction alone is not the most useful information with respect to the business goal. This is because certain elevators may be permanently more prone to

problems than others. For instance, the owner of the elevator may be reluctant in investing in the elevator spare parts which means that there is no use in performing preventive maintenance just because predicted rate of malfunctions is elevated. Better approach is to distinguish the elevators whose condition is detected to have worsened rapidly, as these elevators are most likely to benefit from prompt attention. A slight but meaningful difference to make, which is reflected in the analytics goal. The analytics goal may seem harsh, as it means seeking easy victories and dismissing those elevators that don't benefit from preventive maintenance visits. Nevertheless, it is imperative given the business goal. Another rationale for the analytics goal is that it offers potentially a high yield in comparison to the implementation effort. Namely, it would be straightforward to schedule preventive maintenance visits by taking the prediction into account (as opposed to a simple calendar based scheduling) and collect the benefits with a minor change to the business process itself.

1.3 Aim of the thesis

Aim of the thesis is to extract insights from maintenance records and condition monitoring data, especially with respect to the defined analytics goal. Special interest is in decreasing the overall rate of door malfunctions as a whole by scheduling preventive maintenance visits based on recommendations from a prognostic model leveraging condition monitoring data (as opposed to current calendar based scheduling).

Survival analysis methods, which are widely used in the medical field, are applied for evaluating the so-called survival of elevators. These techniques extract information from partial observations (i.e. so-called censored data). This is useful, as door malfunctions are a rare occasion - most of the observation periods end not by a malfunction, but by a preventive maintenance visit (which is assumed to restore door condition to as good as new). In this context, survival means that elevator doesn't develop a door malfunction. A certain non-parametric statistical estimator is first used on maintenance records alone in order to compare the survival of average elevator to a freshly maintained elevator. This gives an idea how maintenance visit provides value in average, as well as how the rate of door malfunctions generally develops after a maintenance visit. Then, a simple prognostic model is trained with condition monitoring data. Now, this model can

be inferenced for a survival estimate using condition monitoring data . Performance of the model is evaluated via cross-validation, and its shortcomings are discussed. Next, a more complex neural network model is trained with the condition monitoring data and its performance is compared with the simple model. Finally, a method is put forward for leveraging skill of any prognostic model for scheduling preventive maintenance visits based on condition of the elevator. The effectiveness of the method (for reaching the business goal) is examined in retrospect using a real-life dataset, model by model.

1.4 Research questions and research hypothesis

Thus, it has been established that predicting the value of a preventive maintenance visit is a useful analytics goal. Let us proceed by defining the most interesting research questions to be addressed in the remainder of this thesis:

1. How much a preventive maintenance visit improves the door malfunction rate?

2. How does the value of a preventive maintenance visit develop over time?

3. How well certain mathematical models can predict the risk of door malfunction based on condition monitoring data?

Besides answering the research questions, the thesis puts forward a specific method for scheduling preventive maintenance visits that is based on the predicted benefit in terms of improved survival (i.e. time to malfunction) after maintenance. A null hypothesis is that the proposed method is no more beneficial in scheduling the preventive maintenance visits as the current calendar based method. Turning this around, research hypothesis is as follows:

Research hypothesis: Proposed method for scheduling preventive maintenance visits would result in decreased rate of door malfunctions.

1.5 Proposed method

This thesis presents a method for leveraging machine learning in the maintenance business process of the service provider. More precisely, historical data about maintenance visits is used for training a machine learning model whose predictions represent the risk level of developing a door malfunction. Predictions are done on daily basis. Method starts by first forming a real-life dataset of daily condition monitoring data and then labelling each day by using known ground truths extracted from maintenance records. The resulting dataset is then used to train the prognostic model. In this thesis, two models are trained and their prediction performance is evaluated and compared.

Daily predictions can be used for evaluating drift in elevator condition. Elevators having highest drift are thought to benefit most of preventive maintenance. Each day, elevators can be ranked based on the evaluated drift. To examine the research hypothesis, a practical case study is done (in retrospect) to quantify whether maintenance on highly ranked elevators has been historically more beneficial than on average elevator. Case study yields also an estimate on how much the proposed method could reduce the rate of door malfunctions.

1.6 Outline

Thesis is organised as follows. First, in Chapter 2, a literature review on relevant methods is provided. Focus is on understanding what type of methods and methodologies have been used by others to answer similar type of questions. Details of data preparation view are described in Chapter 3, and also the research methodology is described. Research results are presented in Chapter 4, where performance of two prognostic models is reported and the research hypothesis is examined by means of a case study. Finally, conclusions are given in Chapter 5.

2 RELATED WORK

Condition based maintenance (CBM) is discussed in Bousdekis (2018), where literature review on state-of-the-art CBM methods is given. Authors define maintenance as avoid-

ing the equipment breakdown and improving business performance, for example, in terms of productivity, or elimination of malfunctions. They suggest a taxonomy of following three maintenance categories: breakdown maintenance, time-based preventive maintenance, and CBM. In other literature similar categories are referred to as corrective (or reactive) maintenance, preventive maintenance, and predictive maintenance (Wu 2007). Advantage of time-based preventive maintenance over breakdown maintenance is that it improves serving time of the equipment by eliminating some of the breakdowns proactively. However, the time-based preventive maintenance policy does not take the condition of the equipment into account, as equipment is maintained at fixed time interval. CBM (a.k.a. predictive maintenance) policy is enhancing this as maintenance time interval is affected also by the perceived condition of the equipment obtained via regular inspections referred to as condition monitoring. Condition monitoring can be based either on on-site inspections or online harvesting of sensor data or both. A useful separation of concerns is made by authors where a *prognostic model* is developed via condition monitoring and the model is applied on real-time data streams to produce on-the-fly predictions. Then a separate *decision support method* forms the final CBM policy using a custom cost function that takes preferences and choices of the human decision maker into account.

Different prognostic models have been used in literature to model the degradation process of various systems. The aim of these models is to predict system's marginal residual life (MRL) distribution using system covariates obtained via condition monitoring. New predictions are done at each inspection time to be consumed by the decision support system. Multi-state degradation models have been proposed where decision support method is implemented as Markov decision process. However, the majority of the models assume a continuous degradation process where the degradation process is described using, for example, a stochastic model such as Gamma process (Castro 2012).

Before going deeper into the CBM specific models, let us next discuss the more general concepts that go under the name *survival analysis*. Survival analysis deals with set of methods for analysing the time until the occurrence of an event of interest. These techniques have been mostly used in clinical studies to estimate efficacy of medical treatment with time of death as the most common event of interest. For instance, given a set

of patients where some receive an experimental treatment, and others don't, survival analysis can be used to determine (and quantify) whether the treatment had statistically significant effect to lifetimes of the patients. What makes these methods particularly interesting is that they extract information from *censored observations*. For instance, some patients may drop out from the clinical study or may be still alive when study ends. Such patient, not experiencing the event of interest during the observation period, is said to be *right censored*. (There exists also left censored observations, where event of interest has happened before start of observation period, such as emergence of child's first tooth.) Now, in the context of elevators, vast majority of elevators never experience the event of interest (e.g. door malfunction). However, there is value in knowing that an elevator went free of malfunctions until the point of censoring, and the survival analysis techniques are designed to take advantage of that information. The censoring process is said to be *non-informative* if censoring times and event times are independent of each other, meaning that censoring does not leak information about subject's survival.

2.1 Survival functions

Survival function *S(t)* is a function that gives the probability that event of interest (e.g. death or failure) has not yet happened at time *t*. Survival function, which is also sometimes referred to as survivor function (Armitage 2002) or as reliability function (Wolstenholme 1999), is derived from cumulative distribution function *F(t)* using the following formula where *T* is a continuous random variable and *f* denotes its probability density function:

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

The plot of *S(t)* against t is called survival curve.

Survival function lends itself well for estimating lifetime distributions. For instance, probability density function for remaining lifetime, given that subject has survived until a moment t_0, can be written now as follows:

$$P(T \leq t_0 + t \mid T > t_0) = \frac{F(t_0 + t) - F(t_0)}{S(t_0)} \tag{2}$$

Taking derivative with respect to time yields the probability density function of the remaining lifetime at t_0:

$$f_{t_0}(t) = \frac{f(t_0 + t)}{S(t_0)} \tag{3}$$

It is sometimes convenient to use so-called hazard function that gives the hazard rate (i.e. death rate) as a function of time t. Hazard rate gives the probability of subject's death at specific moment in time given that the subject has survived until that moment. Hazard function *h(t)* is derived from survival function and probability density function as follows:

$$h(t) = \lim_{dt \to 0} = \frac{P(t \leq T < t + dt))}{dt \cdot S(t)} = \frac{f(t)}{S(t)} \tag{4}$$

Finally, a cumulative hazard function *H(t)* is defined as integral of the hazard function and its relation to survival function is:

$$H(t) = \int_{-\infty}^{t} h(u)du = -\ln(1 - F(t)) = -\ln(S(t)) \tag{5}$$

Different statistical functions have been used to model the survival function. These can be categorised as non-parametric, semi-parametric and parametric survival functions. Let us proceed by discussing the most relevant survival functions used in the field of survival analysis.

2.2 Kaplan-Meier estimator

Kaplan-Meier estimator is a non-parametric estimator that was first introduced in Kaplan (1958) - a fundamental paper that has been cited tens of thousands of times

since its introduction more than 60 years ago. Kaplan-Meier estimator is built from set of observations and is used to estimate the survival function of the observed entities. Authors define the event of interest as death and previous occurrence of some other event (i.e. right censoring) as loss. The estimator of the survival function is given by the following discrete function:

$$\hat{S(t)} = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{6}$$

Where product is taken over all discrete timesteps t_i where at least one death happened (including timestep *t)*. d_i is the number of deaths at timestep t_i and n_i denotes the number of individuals at risk (individuals that have not yet experienced either death or loss at timestep t_i). This formula produces a monotonically decreasing step function. The value of the estimator function between sampled observations is assumed to be constant. Important characteristic of the Kaplan-Meier estimator is that it leverages also right censored observations (via number of known survivors).

If covariates of the observed subjects are known, such as the genotypes of the patients, it is possible to calculate separate, group-wise estimators for comparison. With large number of patients, it is then possible to visualise (and quantify) how much the different genotypes affect the corresponding survival curves of patients and whether the effect is statistically significant. The model is nevertheless limited in its ability to estimate survival based on covariates due to its non-parametricity. For instance, it can't be used for predicting survival of an individual subject.

2.3 Cox proportional hazards model

The Cox proportional hazards (PH) model (Cox 1972) is a semi-parametric model for approximating survival. Key assumption of the model is that effect of each covariate to subject's survival is multiplicative with respect to the hazard rate. In other words, change in single covariate's value (others remaining constant) lowers (or raises) the hazard rates with a similar amount throughout the time axis. The rate of change is de-

termined by estimated coefficients of the model. Formally, hazard function now takes the form:

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \cdots + \beta_p X_p) = h_0(t) \exp(\boldsymbol{\beta^T X}) \quad (7)$$

Where X is a column vector of the covariates of the subject and β is a column vector of the corresponding coefficients, and $h_0(t)$ is a *baseline hazard function* that is applicable when all covariates are set to zero. The model makes no assumption on the form of the baseline hazard function. If a known distribution is assumed for the baseline hazard function, then resulting model is referred to as parametric PH model. Interestingly, if Weibull hazard function is assumed as the baseline, then the model satisfies requirements of an accelerated failure time model.

2.3.1 Hazard ratio

Thus, hazard ratio is the ratio between individual's hazard function and the baseline hazard function. The model has the convenient property that natural logarithm of hazard ratio $HR(X)$ is a linear function of the covariates:

$$\ln(HR(X)) = \ln\left\{\frac{h(t,X)}{h_0(t)}\right\} = \ln\left\{\frac{h_0(t)\exp(\beta X)}{h_0(t)}\right\} = \beta_1 X_1 + \cdots + \beta_p X_p \quad (8)$$

Solving the linear regression equation yields the coefficients (i.e. proportional hazards). Each estimated coefficient β_i represents an expected change of log of the hazard ratio per unit of change in the corresponding covariate X_i (assuming that all other covariates remain constant). This can be seen easily by writing:

$$HR(X) = \prod_{i=1}^{p} \exp(\beta_i)^{X_i} \quad (9)$$

For instance, if covariate X_i is a dichotomic predictor (e.g. smoking vs. non-smoking), then $\exp(\beta_i)$ yields the its contribution to the hazard ratio. If the contribution is larger than one, then the predictor increases the risk of death. If it is smaller than one, then the predictor is a protective one and promotes survival. In other words, sign of the coeffi-

cient determines whether the corresponding predictor is increasing or decreasing the risk of death. Values close to zero mean that the predictor is not significant. For a continuous predictor (e.g. elapsed days since last preventive maintenance visit to an elevator) $\exp(\beta_i)$ yields the change in hazard ratio per single unit change in the predictor.

Recalling Equation 5, any survival curve *S(t, X)* may be written as:

$$S(t, X) = \exp(-H(t, X)) = \exp(-H_0(t) HR(X)) = S_0(t)^{HR(X)} \quad (10)$$

Where $S_0(t)$ is baseline survival function and *HR(X)* is hazard ratio. This formula is useful when $S_0(t)$ is available and Cox PH model has been fitted for calculating *HR(X)*. The baseline survival function may have been obtained using a non-parametric estimator, such as Kaplan-Meier, or by using a parametric version of the Cox PH model and fitting a known distribution to the observations.

2.3.2 Partial likelihood

Like Kaplan-Meier estimator, the Cox PH model can be fitted from set of observations, while leveraging right censored data. Censoring is handled by a concept of partial likelihood (Cox 1975). Model is fitted to the observations by producing (partial) maximum likelihood estimates (MLE) for the model coefficients. Fitting is typically done using a statistical software package, such as Lifelines (2019) which gives also confidence intervals for the risk ratios.

Key idea in partial likelihood method is to look at death times (i.e. timesteps when at least one death occurred). Let $\lambda_1 < \ldots < \lambda_K$ represent the distinct death times. Let X_i denote the death or loss time of an individual, and Z_i its covariates. Let *R* denote the set of individuals at risk at time t (i.e. individuals not having yet experienced either death or loss). Partial likelihood is the conditional probability of a death of an individual belonging to the set *R* over the death times. At each distinct death time X_j, the contribution to the partial likelihood is:

$$L_j(\beta) = \frac{h(X_j \mid Z_j)}{\sum_{l \in R(X_j)} h(X_j \mid Z_l)} \tag{11}$$

Recalling the hazard function definition for the Cox PH model, the full partial likelihood L over all death times may be written as:

$$L(\beta) = \prod_{j=1}^{K} \frac{h_0(X_j)\exp(\beta Z_j)}{\sum_{l \in R(X_j)} h_0(X_j)\exp(\beta Z_l)} = \prod_{j=1}^{K} \frac{\exp(\beta Z_j)}{\sum_{l \in R(X_j)} \exp(\beta Z_l)} \tag{12}$$

This is the partial likelihood defined in Cox (1975). Note that no assumptions on underlying distribution is made as the baseline hazard function has been cancelled out. Likelihood is a function of observations and model coefficients. The maximum likelihood estimates for the coefficients can be computed by solving the following system of equations:

$$\frac{\partial \log(L)}{\partial \beta_j} = 0, j = 1, 2, \ldots, N \tag{13}$$

The MLE method is known to be asymptomatically consistent, meaning that when number of deaths increases, the estimates are converging to the real values (Armitage 2002).

Cox proportional hazards model is still today the most widely used model in survival analysis. It is safe and proven method to use as it makes no assumptions about the distribution of survival times (Schober 2018). However, parametric models may at times have an advantage over Cox PH model, since they can estimate actual survival curves.

2.4 Parametric survival functions

Different common statistical distributions have been applied for estimating the survival functions of humans or serving time of machines, such as exponential, Gamma, and Weibull distributions (Wolstenholme 1999). Each of these distributions are further de-

fined by their parameters. Exponential distribution has single parameter lambda which is the number of events per unit time (i.e. death rate in the context of survival analysis). Exponential distribution would estimate the survival function perfectly if the probability of death would be constant in each point of time. Obviously, this is not very applicable assumption with respect to condition based maintenance where survival function is assumed to depend on system covariates. Weibull distribution extends the exponential distribution by allowing hazard rate to increase (or decrease) over time. Much more realistic assumption for mechanical equipment, which is why Weibull hazard function is widely used in prognostic modelling. Due to its importance in the field of reliability engineering, let us examine the properties of the Weibull distribution a little closer.

2.4.1 Weibull distribution

The probability density function of a Weibull distributed random variable in its most general form is (Weibull 1951):

$$f(t) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}} \tag{14}$$

Where $f(t) \geq 0$, $\beta > 0$, $\eta > 0$, $\infty < \gamma < \infty$

This is referred to as three-parameter Weibull distribution expression. The three parameters are shape parameter (β), scale parameter (η), and location parameter (γ). When location parameter is not used it is set to zero and the equation is reduced to its most common two-parameter form:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{15}$$

Which in fact is a product of hazard function and survival function, these being:

$$h(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} \tag{16}$$

$$S(t) = e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{17}$$

Following from Equation 5 and Equation 17, cumulative hazard function is:

$$H(t) = \left(\frac{t}{\eta}\right)^{\beta} \tag{18}$$

Effect of the shape and scale parameters to probability density plot is illustrated in Figure 1.

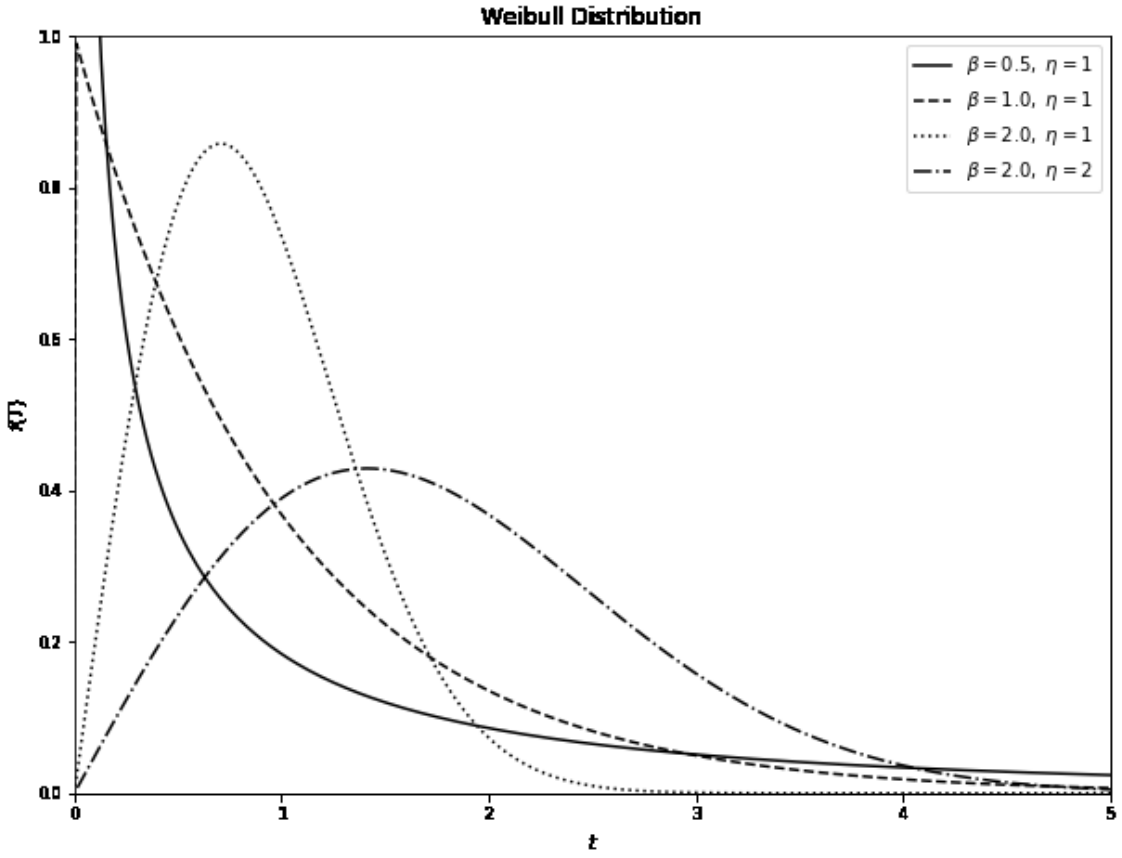

Figure 1. Weibull distribution by different shape and scale parameters.

For $\beta = 1$ the Weibull probability density function reduces to that of the exponential distribution. This constitutes a point of equilibrium where hazard rate is constant. For $\beta < 1$, the hazard rate decreases over time and for $\beta > 1$ it increases over time, as can be seen from Equation 16. For $\beta > 1$ the probability density function has an inflexion point at $(e^{1/\beta} - 1) / e^{1/\beta}$. Increasing the scale parameter, while keeping the shape parameter constant, stretches the distribution further down the time axis. Shape of the distribution will stay similar. The area under the curve must remain the same and therefore stretching causes the peak of the distribution to move lower.

Another visualisation is provided in Figure 2 where Weibull hazard rates, given by Equation 16, are plotted.

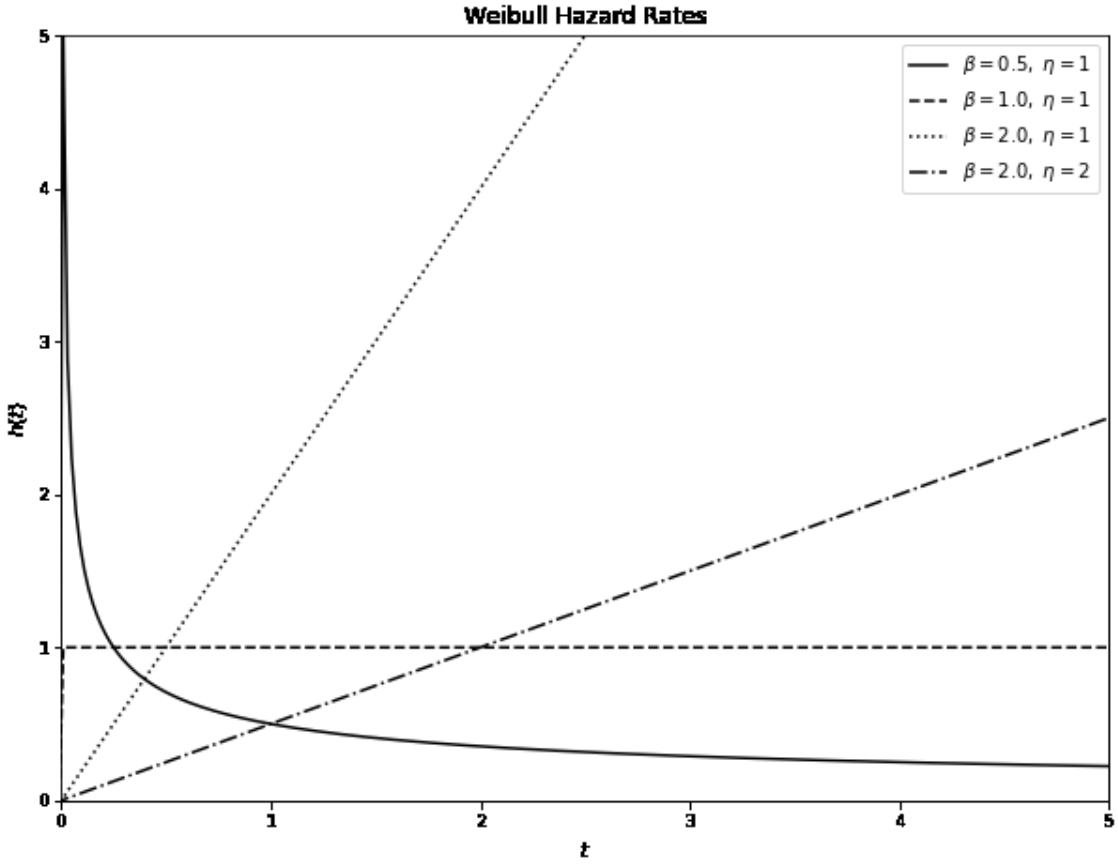

Figure 2. Weibull hazard rates with different shape and scale parameters.

Figure 2 shows how shape parameter controls whether hazards rate increases or decreases over time and how scale parameter controls the rate at which the hazard function changes. With $\beta = 2$ the function increases nearly linearly. With $\beta > 2$ the hazard func-

tion grows more like an exponential function and with $1 < \beta < 2$ the curve resembles a logarithmic function. Figure 3 illustrates how shape and scale parameters affect the survival curve that can be obtained via plotting the Weibull survival function (Equation 17).

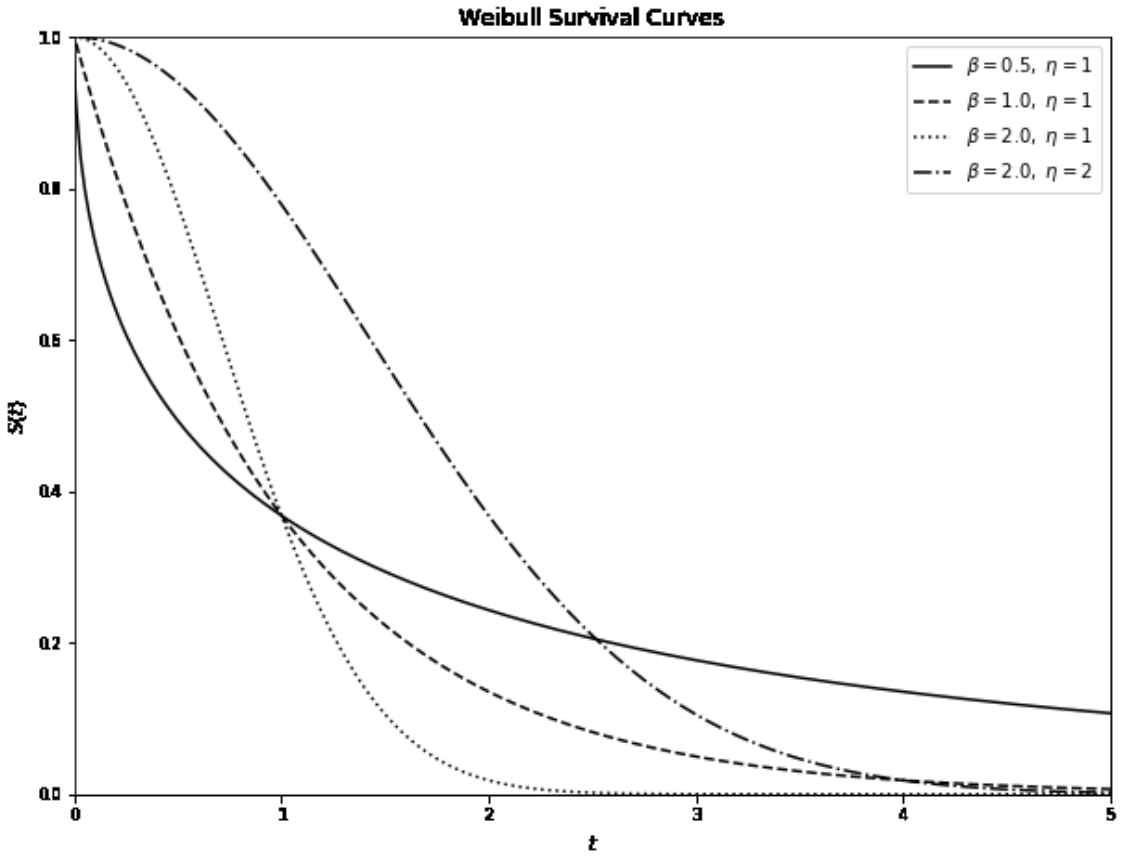

Figure 3. Weibull survival curves.

2.4.2 Weibull plot

When using parametric models, one must be cautious that the used distribution is representative of the underlying phenomenon that is to be modelled. A technique called probability plotting can be used to ensure this. Idea is to plot estimates of cumulative distribution function over time using log-log scale. Assumption that data follows Weibull distribution can be validated using a Weibull plot (Nelson 2004). For Weibull distributed random variable:

$$\ln(-\ln(S(t))) = \ln\left(H(t)\right) = \ln\left(\frac{t}{\eta}\right)^{\beta} = \beta(\ln(t) - \ln(\eta)) \qquad (19)$$

This property allows a visual examination of the appropriateness of the Weibull distribution for the data by plotting *ln (-ln(KM(t))* vs. *ln(t)*, where *KM(t)* is the Kaplan-Meier survival estimate. If the plot implies a linear function could be used to approximate the observations, then it can be assumed that Weibull distribution is applicable. Weibull plot can also be used for fitting the Weibull parameters using linear regression. Least squares fit of a line to the observations produces estimates for the shape and scale parameters (assuming a two-parameter Weibull distribution). Note also, that plotting two survival curves for which proportional hazard assumption is valid produces two parallel lines that are apart by *ln(HR)*:

$$\ln(-\ln(S(t))) - \ln(-\ln(S_0(t))) = \ln\left(\frac{H_0(t)HR}{H_0(t)}\right) = \ln(HR) \quad (20)$$

2.4.3 Parametric likelihood

In Section 2.3.2, it was discussed how partial likelihood handles right censored data when calculating the coefficients for the Cox PH Model using the MLE method. For a parametric model, *parametric likelihood* can be leveraged to handle censored observations. Assuming number of observations to be *n*, then likelihood can be written:

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta) \quad (21)$$

Where θ are the parameters of the model. If a subject experienced death, then its contribution is:

$$L_i(\theta) = f(t_i) = h(t_i)S(t_i) \quad (22)$$

If subject experienced a loss, then its contribution is simply:

$$L_i(\theta) = S(t_i) \quad (23)$$

Thus, likelihood can be written as:

$$L_i(\theta) = \prod_{i=1}^{n} h(t_i)^{d_i} S(t_i) \quad (24)$$

Where d_i is 1 for death and 0 for loss. Taking natural logarithm on both sides, and applying cumulative hazard function from Equation 5, gives the general parametric log-likelihood formula for right censored observations:

$$\ln L_i(\theta) = \sum_{i=1}^{n} \{d_i \ln h(t_i) - H(t_i)\} \tag{25}$$

Again, parameters of the distribution can be estimated by applying the MLE method.

2.5 Concordance index

Harrell's concordance index (C-index) is the most used metric when evaluating performance of survival models. It was developed to evaluate prognostic models fitted with censored data (Harrell 1982). C-index measures how well the model preserves pairwise relative rankings of observations. More specifically, it is defined as ratio of correctly ordered (i.e. concordant) pairs to all comparable pairs. Pair is comparable if the lower of the observation times is for death or the observation times are equal and one of them is for death and the other is for loss. Its interpretation is similar to the area under ROC curve (AUC), meaning that value 0.5 indicates that model has no skill at all, and value 1 indicates a perfect model. The metric is calculated as follows. Let observations be a set of tuples (y_i, d_i) where y_i denotes time to event of interest and d_i denotes type of the observation (1 for death 0 for loss). Assume a prognostic model that can quantify survival from covariates, and the predictions (denoted as p_i) are comparable with each other. Now, a pair of observations $\{(y_i, d_i), (y_j, d_j)\}$ is said to be concordant if $y_j > y_i$ and $p_j > p_i$. C-index yields ratio of such concordant pairs to all comparable pairs. Ties ($p_j = p_i$) can be assumed to be broken randomly. Formally this can be expressed as:

$$c = \frac{1}{n} \sum_{(y_i, y_j) \in C}^{n} I[(y_j \geq y_i, d_i = 1), p_j > p_i] \tag{26}$$

Where C is a set of comparable pairs and indicator I returns 1 for concordant pair and 0 for discordant pair.

2.6 Neural networks

The Cox PH model has its shortcomings. One downside is that interactions between covariates are not accounted for inherently by the model. For instance, being a smoker could double the risk of a cardiac arrest, and high blood pressure might do the same. The model would merely implicate a quadruple risk for an individual with both condi-

tions. In real-life the risk could be higher, due to the predictors amplifying each other. One could apply manual feature engineering to remedy this problem or use a more complex model that is able account for feature interactions inherently. See page 28 in Martinsson (2016) for a discussion of different approaches that have been developed to remedy the shortcomings of the Cox PH model with this respect. Yet another nuisance is that Cox PH model only yields hazard ratios for predictors. Generally, it is more useful to know the absolute risk level instead. Thirdly, unlike recurrent neural networks, which can extract temporal relationships of condition monitoring data at different inspection steps, traditional survival models don't inherently have any notion of time-variate covariates. This drawback too needs to be tackled with feature engineering, which adds complexity. No doubt that feature engineering is a discipline on its own and much progress has been done in that area. However, it is also a laborious process, and sometimes time is of essence. For these reasons, let us look beyond the traditional survival models in this thesis. Neural networks are known to be able to approximate complex mapping functions so why not also the one between covariates and survival curves. In addition, neural networks, unlike popular decision-tree methods, can inherently model also temporal interactions via addition of recurrent layers. A lot of papers on applicability of neural networks on survival analysis has been published recently. A concise literature review on these is provided in Sub-section 2.6.2.

2.6.1 Advent of neural networks

Artificial neural network (ANN) as a concept is sometimes credited to a paper published in 1940's McCulloch (1943), where authors modelled a simple network of artificial neurons, inspired how neurons in human brain might work, using electrical circuits. Even more often, invention of the concept is credited to psychologist Frank Rosenblatt who described the Perceptron, a model how a human brain learns to recognise objects (Rosenblatt 1958). However, a fundamental practical work was introduced in the field of electrical engineering by Professor Bernard Widrow and his graduate student Ted Hoff (Widrow 1960) who implemented an adaptive signal processing circuitry. It was named ADALINE, for "adaptive linear neuron" and it implemented a binary classifier for pattern recognition of digital inputs. Training of the neuron was implemented by using an 4x4 array of toggle switches to feed input patterns and a single toggle switch to signal

the desired binary output. Adapting of the weights was governed by method called steepest descent and a loss function that was later named as least mean square (LMS). Weights corresponded to potentiometer settings that were adjusted manually. Figure 4 is from Widrow (2005) where Professor Widrow casually tells the story 45 years later.

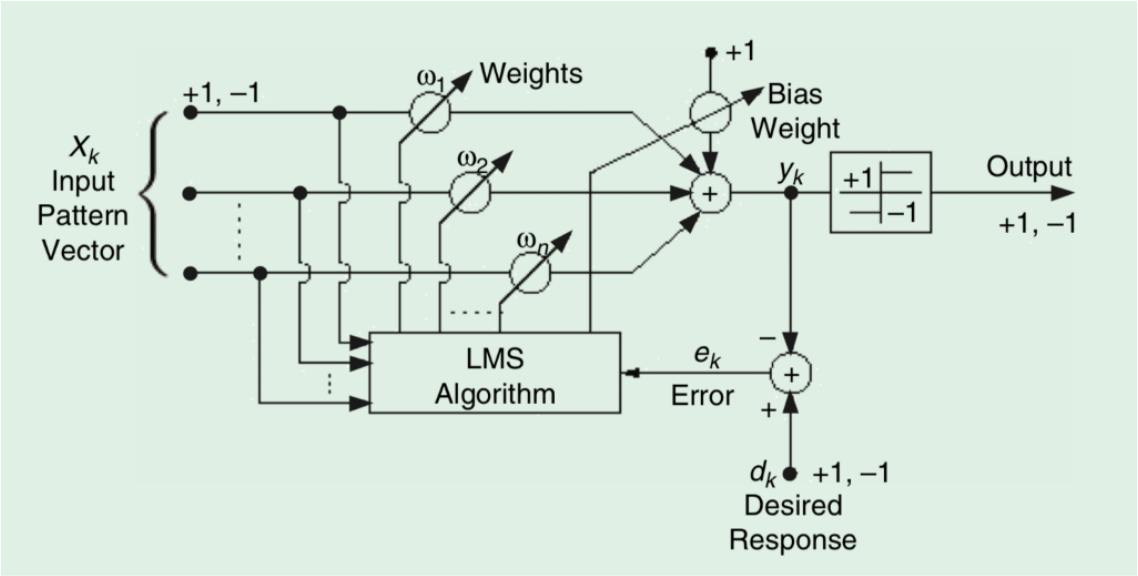

Figure 4. ADALINE circuitry (Widrow 2005).

ADALINE was followed by MADALINE, which was a three-layer fully connected feed-forward neural network based on ADALINE circuitry. It was the first commercial application of a neural network, implementing an adaptive filter for echo cancellation on telephone lines. Since then the number of applications of neural networks have exploded, but still pattern recognition is where they excel.

Much of the early foundation still holds, while new concepts have submerged over the years such as backpropagation, stochastic gradient descent, recurrent neural networks (RNN). In the remainder of the paper, a term neural network is used as an umbrella term to refer to any kind of algorithm building on the concept of neurons. More specific terms, such as RNN, are used to refer to more specific architectures of neural networks.

2.6.2 Neural networks for survival analysis

When it comes to using neural networks for survival analysis, research has been done mostly in the medical field. In the following, let us review some of the most relevant

papers. Given the characteristics of CBM, interest lies particularly on methods where neural network is trained with right censored data and a type of RNN is used.

A model consisting of hierarchical system of neural networks for predicting survival of AIDS patients is put forward in Ohno-Machado (1996). Trained model can be queried for six discrete time points on the survival curve (1, 2, 3, 4, 5, and 6 years), meaning essentially the problem of assessing the survival curve is translated to a binary classification problem with six different prediction horizons. Survival curve begins at the time of AIDS diagnosis. Censored data is discarded in the study. System is computationally equivalent to a recurrent neural network, where covariates are static in each time step. This is because no input data are fed into the model yearly, but predictions are based only on baseline covariates (e.g. biological markers for decease progression) obtained by the date of the diagnosis. The covariates include demographic and socioeconomic explanatory variables, biological markers, clinical findings, and medications. Authors provide comparison between the developed model and a simpler method where six isolated feed-forward neural networks are used to obtain predictions for the six time points on the survival curve. All networks are trained by backpropagation. Area under receive operating characteristics (AUC) is used as the key metric. Proposed method is shown to be superior in all six prediction horizons. In addition, the number of non-monotonic intervals seems to be far worse with the simpler method. Yet, also the proposed approach has potential to produce (discrete) survival curves that are not monotonically decreasing.

In Wu (2007) authors define a neural network based prognostic model and a decision support method for predictive maintenance of rotational equipment (i.e. rolling element bearings). Decision support method is a cost function that optimises the expected cost per unit operational time. Authors define a degradation signal based on condition monitoring data from vibration sensors and define a failure threshold. Artificial neural network (ANN) model is used to predict the life percentage of the equipment from the degradation signal and operating time at each inspection time. Then authors use a separate table for correcting the error made by ANN as compared to actual MRL distributions (table is obtained via using ANN on a validation dataset). Corrected MRL predictions are fed to a cost matrix that suggests optimal replacement times for the machinery (i.e. balance between breakdown maintenance and preventive maintenance).

2.6.2.1 WTTE-RNN

In Martinsson (2016), an RNN is used to learn Weibull distributions from right censored observations. The name WTTE-RNN stands for Weibull time to event RNN. Training dataset is a fixed size continuous matrix whose rows correspond to subjects and columns correspond to time steps. Items contain the covariates and ground truth which is a tuple (*t, d*) where *t* denotes the number of timesteps to next event, and *d* denotes the type of event (1 for death and 0 for loss). Trained model can be inferenced with a matrix of similar shape, where number of rows and number of columns can be arbitrary. As output, the model gives a matrix whose rows and columns correspond to the input and items contain estimated Weibull distributions for each subject on the corresponding timestep. Typically, the last column is the interesting one, as it contains the most knowledgeable predictions. Model assumes recurring deaths are possible, but derivative works have submerged already where WTTE-RNN model is applied for survival analysis (Deep-ttf 2018).

Author derives continuous and discrete loss functions for training the model. What makes them interesting is their mathematical soundness. The loss functions are derived from the general parametric log-likelihood formula (Equation 25) for right censored observations. This means that, if the censoring process is non-informative (i.e. censoring times and survival times are independent of each other), and survival times follow Weibull distribution, then the loss function used by this model is the ideal one.

2.6.2.2 RNN-SURV

In Giunchiglia (2018), a deep recurrent neural network model called RNN-SURV is presented. The model can be used to predict a personal risk scores and survival curves for medical patients. A comparative study is described where RNN-SURV superiority over competing models is reported using different well-known medical datasets. As the model uses censored data, authors use C-index as the main performance metric, and report improvement up to 28.4% over state-of-the-art methods. Authors provide also visualisation where obtained survival curves are compared to the average Kaplan-Meier survival curve and show that individual survival curves can be very different than the average survival curve. Like in Ohno-Machado (1996), the problem is translated to a combination of binary classification problems with different time horizons. The model

outputs an array of tuples (S_i, r_i) where each tuple contains estimates in the end of the corresponding time horizon. S_i denotes the estimate of the survival curve and r_i and denotes a risk score that is computed as a linear combination of the S_i estimates. All covariates are static in each timestep, except for the timestep number that is used as one covariate. Model parameters are number of feed-forward layers, number of recurrent layers, number of neurons on each feed-forward layer, and LSTM state size.

Since the network predicts both survival curve and risk score, the authors define also the loss function as a combination of two separate functions. First one is a modified cross-entropy function that accounts for censored data and second one is an upper bound of the negative C-index (Raykar 2008).

2.6.2.3 RankDeepSurv

In Jing (2019), a deep feed-forward neural network is proposed for predicting survival curves of medical patients. A case study with different medical databases is performed and authors show an overall superiority over other survival models, such as Cox PH model. For one dataset, the model performs better than human clinical experts. C-index is used as the metric, and bootstrap method on the test set is used to compute confidence intervals.

The RankDeepSurv ANN is a fully connected feed-forward neural network except for output layer where dropouts are added to tackle overfitting. Number of layers is a parameter of the model and can be changed (e.g. to match the amount of available data). Authors define a loss function specially designed to handle censored data. The loss function is sum of two terms L_1 and L_2. Interpretation of L_1 is that deaths always contribute to the function, while losses contribute only if predicted value is smaller than the observed value. L_2 gives pairwise contributions of the observations to the loss function. Interpretation of L_2 is that pairwise ranking difference of subjects shall be preserved by the ANN (recalling that two losses are not comparable). The loss functions are completely heuristic as no proof is given where they would be derived from any known formula, such as the general parametric log-likelihood for right censored observations (Equation 25). Proof of convergence is however given. Due to formula L_2, size of the

training set is $O(n^2)$, n being the number of observations. This may prove to be impractical.

2.7 Multimodality of the data

Condition monitoring data is sequential data where measurements are done at each inspection step. Such data can be organised in a 3-dimensional matrix, where rows correspond to data sources (e.g. an individual elevator) and columns correspond to inspection times. Third dimension is the number of features, as each item in the matrix contains all measurement values of an elevator at inspection time. Let us refer to this matrix as dynamic feature matrix since it contains the temporal measurements as dynamic features. Now, there exists also static features that remain constant throughout inspection steps. In case of an elevator, such static features may be elevator make and model, or door motor type. These data are stored in a 2-dimensional static feature matrix. Now, the data on elevator is referred to be multimodal, meaning it is a collection of dynamic and static features.

Traditional machine learning models, such as decision trees or feed forward neural networks, are meant to be used with unimodal data. Usually this means that all covariates are assumed to be static. This hasn't nevertheless prevented researchers from using the models with dynamic features. A common approach is to supplement static features obtained from static feature matrix with a slice from the dynamic feature matrix in order to obtain a sample. Complete set of samples is formed by iterating over inspection times. The slice represents a history window which consists of consecutive values of measurements since start of the history window. Number of features seen by the model is proportional to both number of dynamic features and size of the history window. Thus, this approach suffers from the curse of dimensionality. To deal with this problem, another approach is to summarise the dynamic feature matrix (or slice of it) by curating each dynamic feature using either statistics, stochastic model or machine learning model. Curated features can then be fed into the model instead of full history window data. One such approach for extracting characteristics of multivariate time series is given in (Christ 2018) where python package called tsfresh is documented. In Leontjeva (2016)

authors describe another approach where a model, such as RNN, is used for producing the curated features from dynamic features as illustrated in Figure 5.

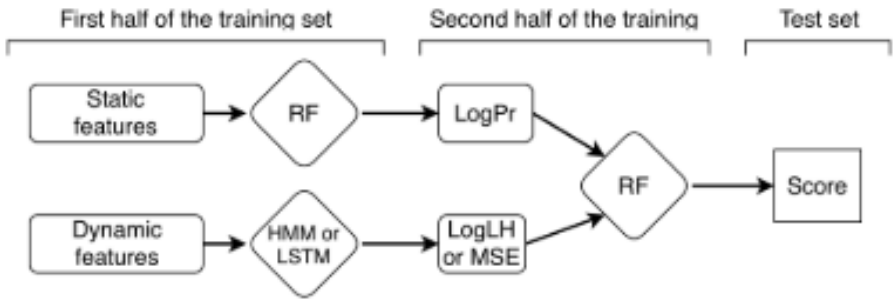

Figure 5. A pipeline that can use RNN for feature transformation (Leontjeva 2016).

Idea is that the first stage model is trained first using first half of the training set. This model can be either a hidden Markov model, or RNN. Then the predictions from the first model are used as new set of features by the second stage model, which in the paper is a Random Forest classifier. In other words, the final model is a pipeline consisting of RNN based feature transformer and RF based classifier. To be more precise, predictions from the RNN are mean squared error of a sample per each class rather than the actual binary prediction. In this respect the approach resembles so-called transformation learning where only last layer of neurons of a trained model is trained again using a secondary dataset.

Recurrent neural networks, on the other hand, assume features to be dynamic. They are inherently seeking temporal interactions between consecutive samples. Therefore, RNN is not able to consider any static features unless they are transformed into a dummy time series data. In other words, the dynamic feature matrix items must be supplemented with the values of static features. Same set of static covariates are now concatenated to dynamic measurement values at each inspection time, which leads to growing the third dimension of the dynamic feature matrix by the number of static features. The supplemented matrix is then used by the model as shown in Figure 6.

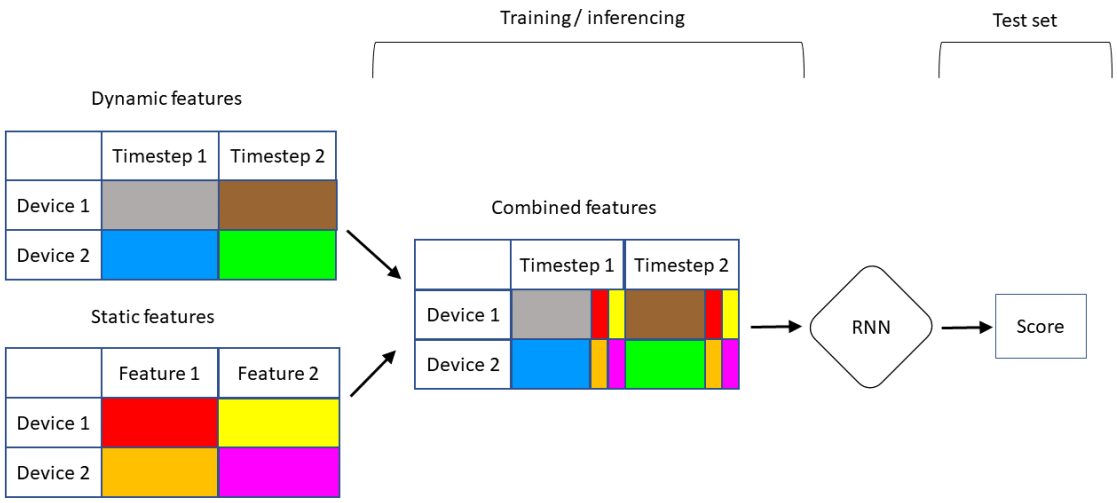

Figure 6. Combining static features and dynamic features by making a dummy time series out of static features.

3 RESEARCH METHODOLOGY

The research questions are analysed by a series of survival analysis exercises. Real-life condition monitoring data on more than 20 thousand elevators is used. Time span of the data is ca. 2 years. Machine data is combined with site visit data, in order to form a labelled dataset. Resulting dataset focuses on door malfunctions, other types of malfunctions are left out. Ratio between non-censored and censored observations is approximately 1:30. Censoring process is assumed to be non-informative, meaning that malfunctions, preventive visits, and censoring times are assumed to be independent of each other.

First, Kaplan-Meier estimator is used to visualise the general effect of a preventive maintenance visit with respect to survival time. Samples are stratified based on whether they are for freshly maintained elevators or not and obtained survival curves are plotted. Then, Weibull plots of the Kaplan-Meier estimates are drawn, in order to assess whether Weibull distribution may be used for estimating survival times, and to see how well proportional hazards assumption holds.

Two prognostic models are fitted with the dataset, and corresponding learning curves are plotted using C-index as the performance metric for comparison. Models can be inferenced to obtain daily risk predictions using historical data of the training set. Cox PH model yields directly hazard ratios, while WTTE-RNN yields Weibull survival curves that too can be ranked by fixing a time horizon. Elevators are assigned daily ranks based on how much the predicted risk has been increased during the ongoing observation period. Finally, highly ranked elevators are compared to all elevators to examine the magnitude of the maintenance effect in retrospect. Null hypothesis is that effect of a preventive maintenance visit is not any different within the high-rank group than it is on the whole population of elevators.

3.1 Dataset

The dataset, that was crafted to fuel the models, consists of a dynamic feature matrix combined with static feature matrix as depicted in Figure 6. The combined feature matrix is a multi-indexed dataset, containing 10 daily condition monitoring daily aggregates for different elevators, and 3 static features as dummy time series. Rows of the dataset correspond to elevators, and columns correspond to days. Data preparation is started by forming daily condition monitoring aggregates, per elevator. For example, number of total door openings was one of the condition monitoring covariates that was aggregated per calendar day (see Table 1 for other covariates). Only elevators for which at least 90% of days had daily condition monitoring aggregates available were accepted to the dataset. Missing days (i.e. inspection steps) were filled using forward fill policy. Then, the obtained dataset was merged with maintenance records to denote where in the past and in the future the next preventive or unplanned visits have occurred. This information gave birth to two more covariates called *DaysSincePreventive* and *DaysSinceUnplanned*. Finally, static features were merged by elevator to form three static covariates, called *NumberOfFloors*, *NumberOfLandingDoors*, and *BuildingType* (a categorical feature, which was still one-hot endoded before training the models). Final dimensions of the dataset seen by the models were 20465 rows, 670 columns, and 27 covariates.

3.1.1 Labelling

Each sample in the dynamic feature matrix represents an observation which is a tuple (t, d) where t denotes number of days to next observation and d denotes the type of observation (1 for death, 0 for loss). In this case, death corresponds to a reported door malfunction. Loss can happen for two reasons. It may be that there is no data available after censoring time. Another reason for a loss can be that preventive maintenance visit has been done before any door malfunction has occurred. It is assumed that doors are then fixed to as good as new condition, which concludes the ongoing observation period and starts a new one. Usually malfunctions are repaired immediately during the same day when they were reported. In a rare case where malfunction stays active more than a single day, all active days are labelled as deaths.

3.2 Training and cross-validation

Learning curves are produced using 3-times repeated group 3-fold cross-validation where the splits are done on elevator basis in order to avoid leaking information between test set and training set. This means that altogether 9 different C-index scores are calculated per training set fraction and average of these observations form each of the points in the learning curve. Same elevator never appears in both training and test sets. This method was selected to ensure that results will represent the true predicting power of the model on unseen data. Validation of the research hypothesis is done using the same method and with full dataset. In other words, the lift curves represent an average of 9 different cross-validation rounds.

3.3 Performance

Models are tuned and compared by C-index as the score, since that is designed to be used with censored data. Downside with C-index is its complexity $O(n^2)$. Therefore, number of observations is truncated to 100 000 by random choice without replacement when C-index is calculated. C-index is not linked to the research hypothesis very well. Hence, additional metrics are used to quantify the effect of preventive maintenance visit within a group of elevators (e.g. high-risk elevators). First, let us define a metric called

lift and define it as a ratio of cumulative hazards of two survival curves. Let us denote $L_{ij}(t) := H_j(t) / H_i(t)$. Lift can be interpreted as the inverse of the risk ratio for an interval t of two elevator populations. For example, consider two survival curves i and j that correspond to two distinct elevator populations. Now, if $H_i(t) < H_j(t)$ risk ratio would be less than 1, which would indicate a decreased risk level being associated to group i as compared to group j (i.e. when comparing their probabilities to survive until time t). The smaller the risk ratio, the greater the lift. Research hypothesis can now be defined formally, as an expectation that for most intervals t, $Lift_A(t) > Lift(t)$, where $Lift_A$ is calculated for high-risk elevators, and $Lift(t)$ for all elevators.

Finally, a metric called *gain* is defined as the ratio of $Lift_A(t)$ to $Lift(t)$. Let us denote $Gain(t) := Lift_A(t) / Lift(t)$. Gain is interpreted as the benefit of using model based scheduling over random scheduling. It yields the expected reduction in the number of unplanned visits as a function of days passed after the maintenance which is a relevant metric considering the business goal which is to decrease the rate of unplanned visits.

4 RESULTS

In Section 1.3, a number of research questions were posed, and a research hypothesis was given. In the following, research questions are answered by leveraging the constructed real-life dataset. This chapter is concluded by examination of the research hypothesis via case study.

4.1 Effect of a preventive maintenance visit

Let us examine how a maintenance visit affects survival curves. See Figure 7 for two survival curves.

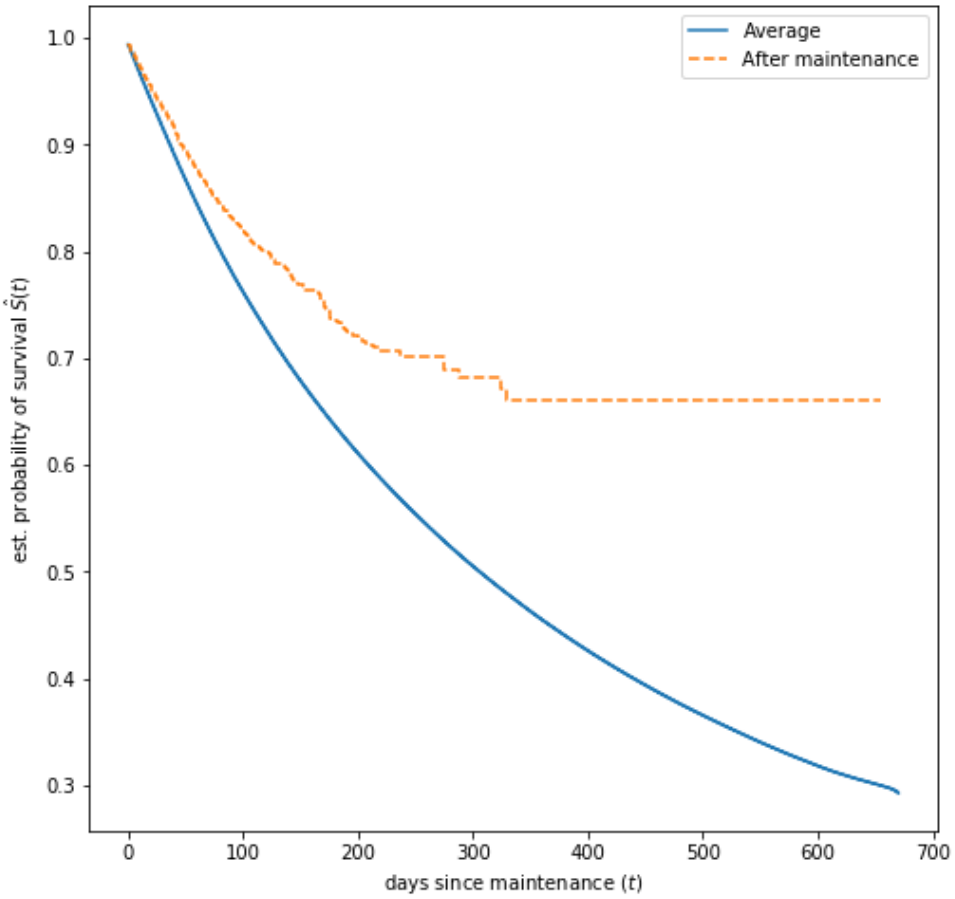

Figure 7. Average survival curve and survival curve after preventive maintenance.

The two curves were obtained by applying Kaplan-Meier estimator. Average survival curve was produced by considering all observations in the dataset while second curve was obtained by filtering observations where planned maintenance visit had been done on previous day.

4.1.1 Weibull analysis

Next, the Weibull plot technique, described in Section 2.4.2, was used to validate whether the survival times can be assumed to follow the Weibull distribution. Figure 8 shows the plot of the average survival curve and the baseline survival curve.

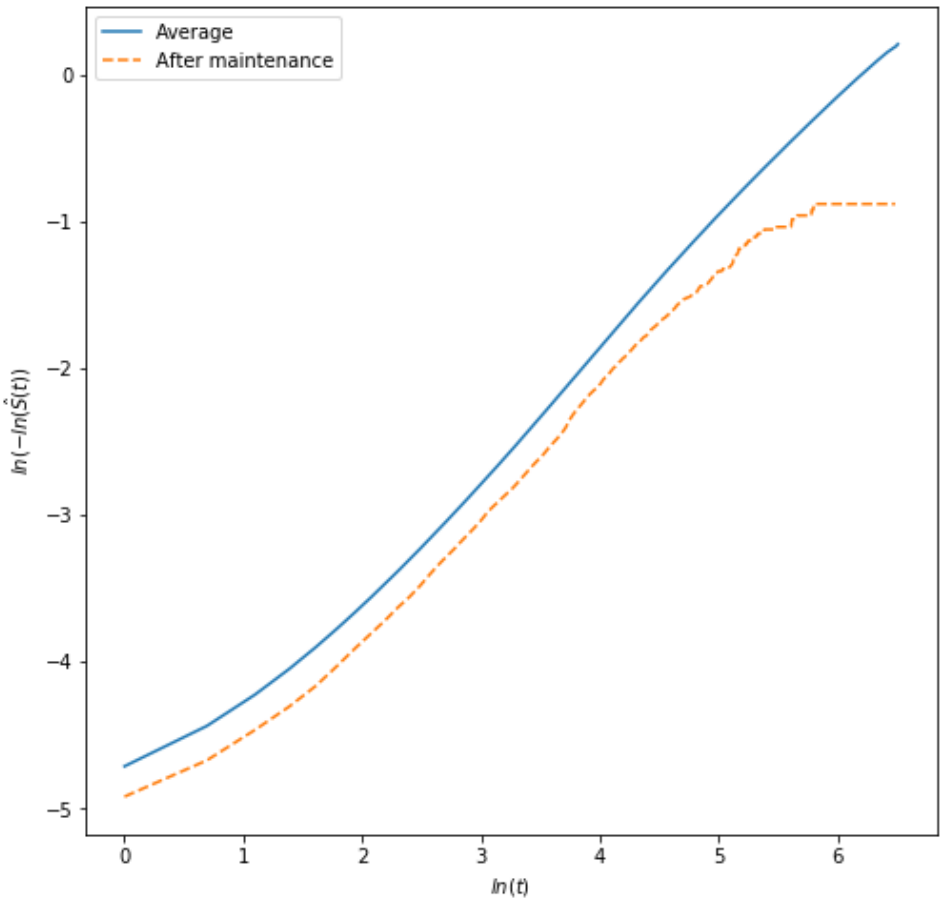

Figure 8. Weibull Plot of the Kaplan-Meier survival curves.

Especially the average survival curve seems to follow the Weibull distribution well, since the plot implies a linear relation between x and y axis. The baseline survival curve follows reasonably well too, so it can be concluded that it makes sense to try out Weibull based survival analysis techniques. Corresponding trendlines that were fitted using the least squares method are shown in Figure 9.

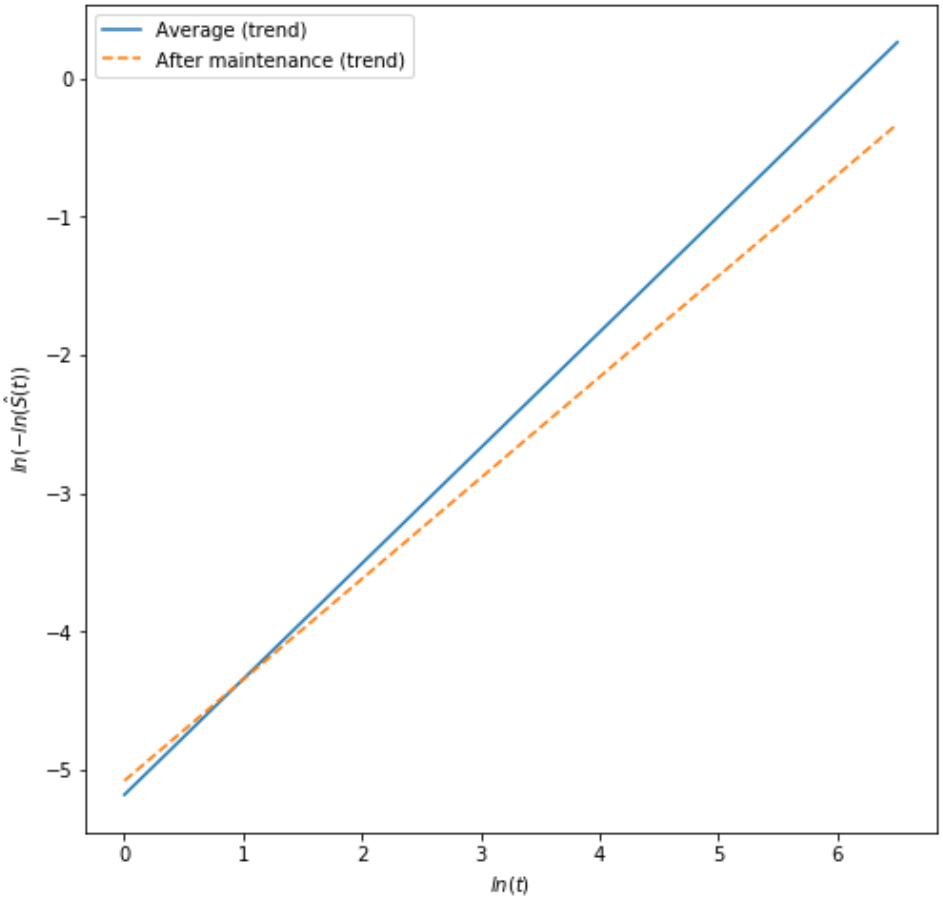

Figure 9. Trendlines for Weibull plotted Kaplan-Meier survival curves.

The trendlines are not parallel, which implies that Cox PH model might not yield the most accurate results. For the Cox PH assumption to hold, the difference of the trendlines should be constant, *ln(HR)*. Now it seems that the real-life hazard ratio is not constant, but proportional to time. Next, a Weibull fit was done with the observations corresponding to the two survival curves using Lifelines (2019). Figure 10 shows different Weibull visualisations for the two populations.

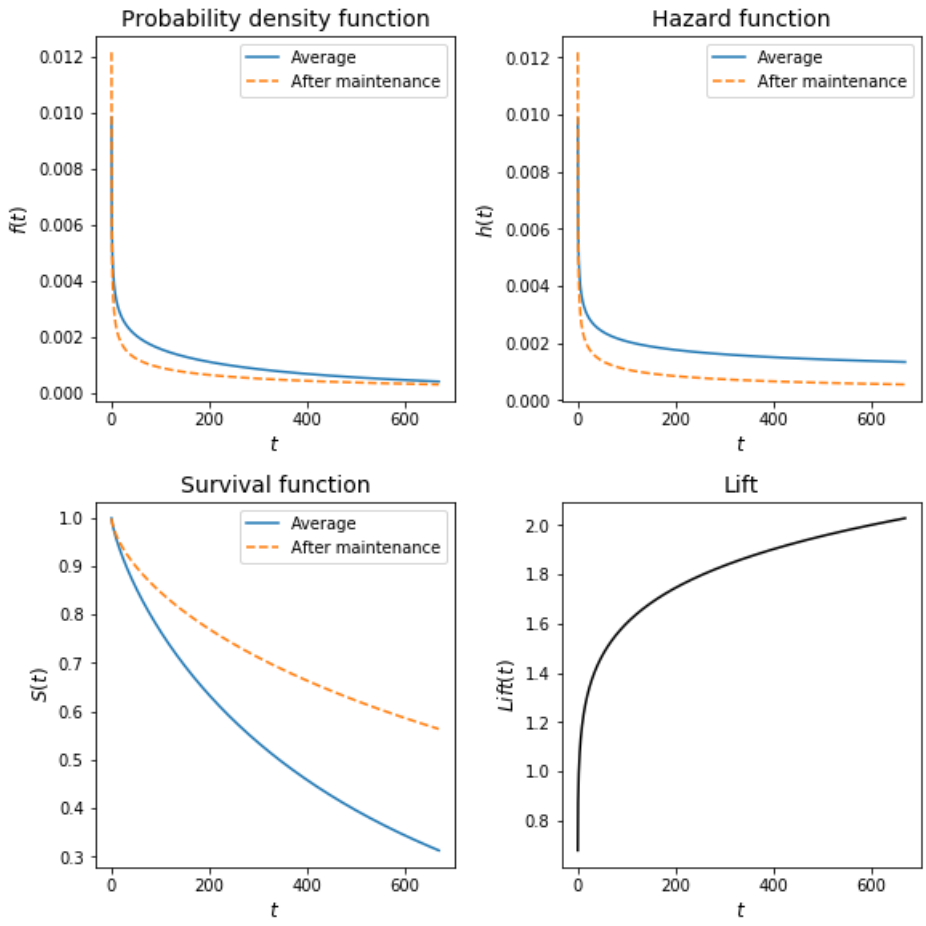

Figure 10. Weibull distributions based on Kaplan-Meier survival curves.

Now that Weibull estimates are known, let us try to answer the research question 1, which is *"How much a preventive maintenance visit improves the door malfunction rate?"*. This is best summarised by top right graph where hazard functions are graphed for the two elevator populations. It shows that maintenance clearly decreases the rate of door malfunctions, and that the effect is quite permanent. Subtracting the means of the two Weibull distributions yields 98 days. However, the bottom right *Lift* curve encodes even more interesting information. Assuming an uninterrupted observation period *t*, *Lift(t)* yields the ratio of unplanned visit rates of an average elevator and a freshly maintained elevator. In other words, *Lift(t)* quantifies, on average, the future benefit of performing preventive maintenance and therefore answers the research question 2, which is *"How does the value of a preventive maintenance visit develop over time?"*. The *Lift* will be leveraged also later, when the research hypothesis is examined.

4.2 Model validation

Before testing for the research hypothesis, let us first examine the Cox PH model and WTTE-RNN model. C-index can be used to evaluate their performance. Evaluations were done by running Python-based Jupyter notebooks on a laptop with 16GB of RAM and Intel core i5-6300U CPU (2.40 GHz).

4.2.1 Cox PH model performance

Cox PH model was fitted with the covariate matrix. More precisely, a Cox PH model with elastic net penalty was used, as standard Cox PH model is not able to cope with the fact that some covariates were highly correlated with each other. Before fitting, all covariates were scaled to same range (from 0 to 10). Actual fitting took 20 minutes. Table 1 shows the obtained hazard ratios per unit for each covariate.

Table 1. Cox PH model hazard ratios per unit of change.

Covariate	Hazard ratio (per unit)
NBR_DOOR_OPERATIONS	1.479
NBR_DOOR_REOP_BY_LIGHTCURTAIN_CUTS	1.391
NBR_DOOR_REOP_BY_OPEN_BUTTON	1.0
NBR_DOOR_REOP_BY_SAFETY_EDGE	1.0
NBR_DOOR_REOP_BY_UNKNOWN_REASON	1.0
NBR_LIGHTCURTAIN_CUTS	1.0
NBR_NUDGING	1.0
NBR_PHOTOCELL_CUTS	1.161
NBR_STARTS_DOWN	1.0
NBR_STARTS_UP	1.0

DaysSinceUnplanned	0.819
DaysSinceMaintenance	1.063
NumberOfFloors	1.321
NumberOfLandingDoors	1.047
BuildingType	[0.929,1.126]

The strongest risk increasing predictors are NBR_DOOR_OPERATIONS, NBR_DOOR_REOP_BY_LIGHTCURTAIN_CUTS, and *NumberOfFloors*. Strongest risk decreasing predictor is *DaysSinceUnplanned* (perhaps if elevator has run for a long time without door malfunctions, it is likely to do so in future also). *BuildingType* expands to so many one-hot encoded columns that let us just make note of the range that these covariates are in and that none of them makes a very strong predictor. Recall that no trends or interactions are provided to the model. It is possible that via feature engineering one could construct better predictors.

Figure 11 shows a box plot of hazard ratios with 95% confidence intervals, variance, and mean. Figure 12 shows corresponding survival curves, where average Kaplan-Meier estimate from Figure 7 is scaled by ratio of the means of hazard ratios obtained via Cox PH model. Compared to stratified Kaplan-Meier estimators and Weibull fit techniques (cf. Figure 7 & Figure 10) the Cox PH model significantly underestimates the maintenance effect, due to the PH assumption being slightly off for the dataset.

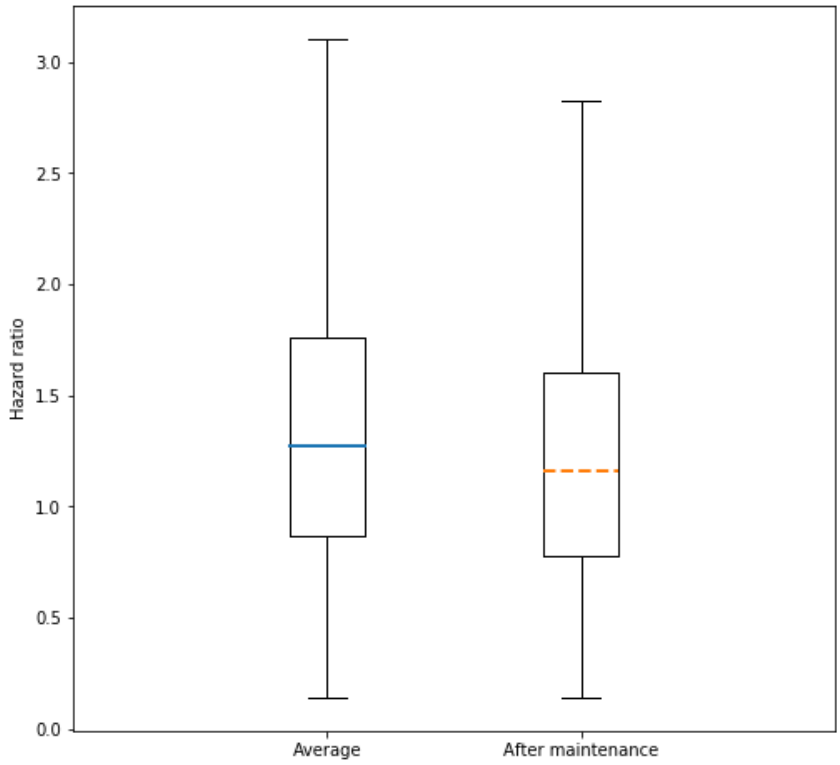

Figure 11. Distribution of hazard ratios.

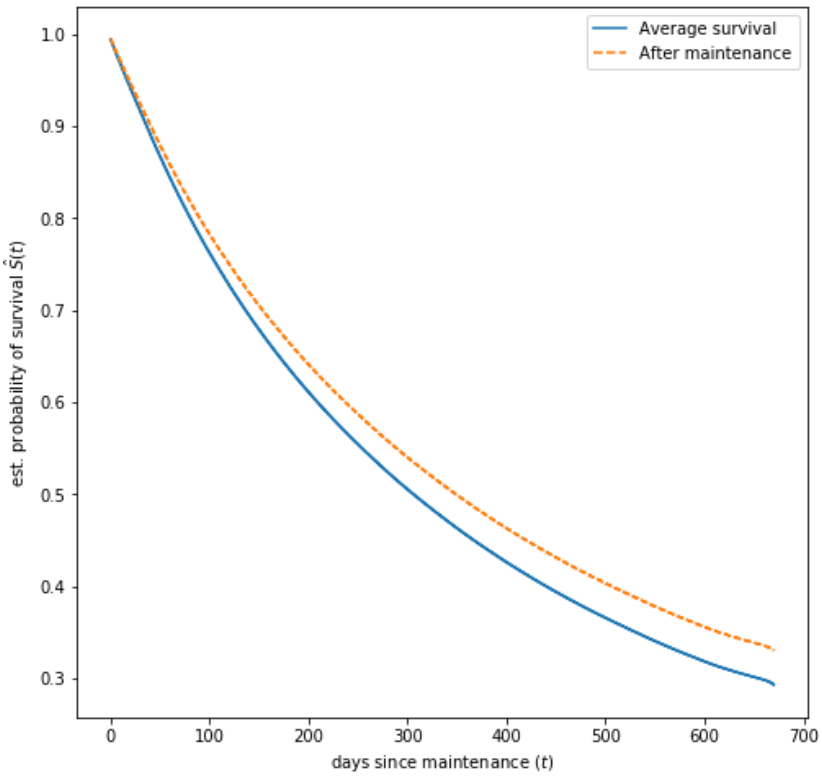

Figure 12. Effect of preventive maintenance by Cox PH model.

Figure 13 shows how C-index develops as function of training set size for the Cox PH model with elastic net penalty. Curves were obtained through 3-times repeated 3-fold group cross-validation. Hence, the model was trained 9 times (3 x 3) per each of the points in x-axis. Graphs shows the means of these scores, as well as the standard deviations as function of training set size.

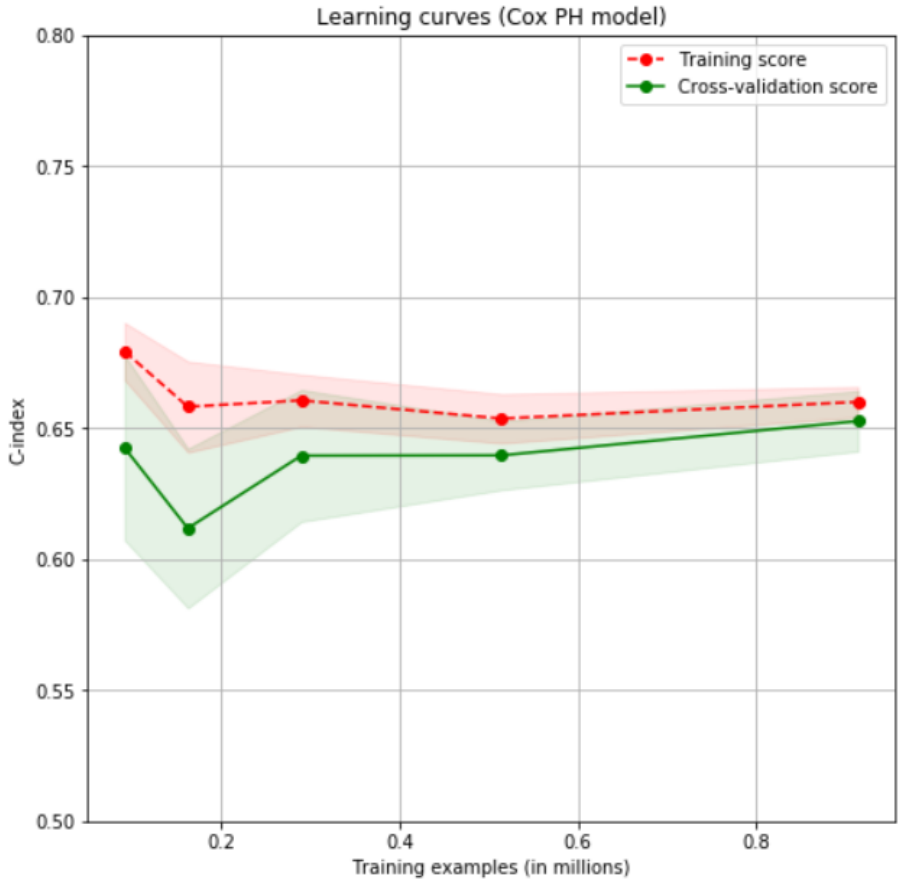

Figure 13. Learning curves for Cox PH model with elastic net penalty.

Graphs indicate that the performance converges quickly and adding more samples does not help much after training set size of one million samples where model is achieves its maximum C-index of 0.65. Early saturation of the learning curve was to be expected as the model is so simple. On the other hand, the model seems stable and overfitting is not an issue.

4.2.2 WTTE-RNN model performance

Let us next build a neural network model based on WTTE-RNN (Martinsson 2016). Author of the model has released the source code in Python WTTE-RNN (2019) and it was used for training the model. Training set consists of a matrix whose rows correspond to elevators, columns correspond to days, and contents are the daily condition monitoring covatiates and ground truth tuples (t, d), where t denotes time to event, and d is 1 for death and 0 for loss. Figure 14 shows the learning curves for the model. Training of the model for largest training set size took 40 minutes. Model consisted of two feed-forward layers, followed by one LSTM layer, one dense layer, and final output layer, and had altogether

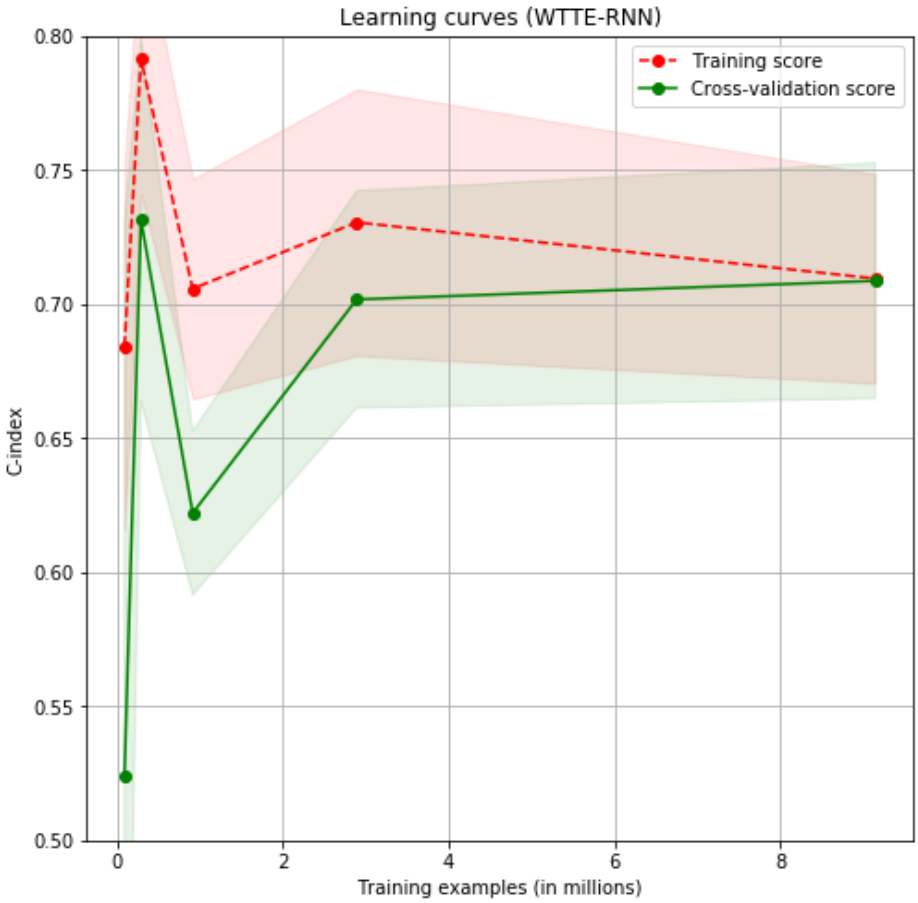

Figure 14. Learning curves for WTTE-RNN.

Learning curves show that the model benefits of adding more samples, and that model generalises well. At around 10 million samples, the performance on test set equals that of the training set C-index being around 0.71, which is about 6 percentage points higher than that of the Cox PH model in saturation.

4.3 Validation of the research hypothesis

The research hypothesis was carved in Chapter 1 and formulated formally in Chapter 3. To validate the hypothesis via case study let us graph metrics $Lift_A(t)$ and $Lift(t)$ for both models.

4.3.1 Lifts from Cox PH model

To calculate lift for Cox PH model, 3-times repeated 3-fold group cross-validation was performed on the dataset to label samples as either high rank or low rank samples. For each round, train set was used to fit Cox PH model with elastic net penalty and fitted model was used to predict risk ratios for the test samples. Then each prediction was normalised by subtracting the baseline risk ratio which was the prediction on the first day of the observation period. This gave the rank of the elevator, which can be interpreted as potential to benefit from a maintenance visit. Top 15% of the ranks were deemed as high rank and the rest as low rank. This threshold was selected ad hoc, because it provided enough elevators in high rank group that had been maintained (computing survival curves for a group that has under 20 members is not reliable, and used math package gave a warning about that). Weibull fit was done on all rounds, and corresponding survival curves and lift curves were calculated. Figure 15 shows average of the survival curves for high rank elevators (group A) and all elevators.

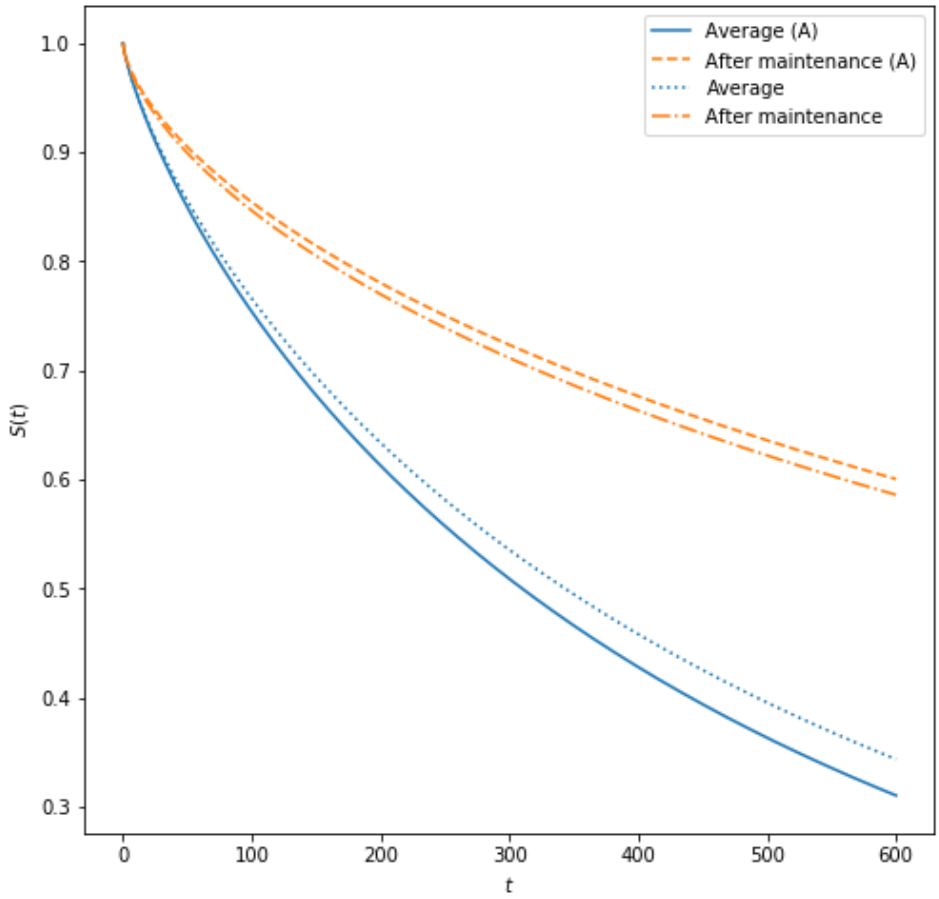

Figure 15. Survival curve estimates in the case study with Cox PH model.

Research hypothesis is validated by examination of the graphs $Lift_A(t)$ and $Lift(t)$ that are shown in Figure 16 as well their relation referred to as $Gain(t) := Lift_A(t) / Lift(t)$. Left y-axis is for the lifts, while right y-axis is for the gain.

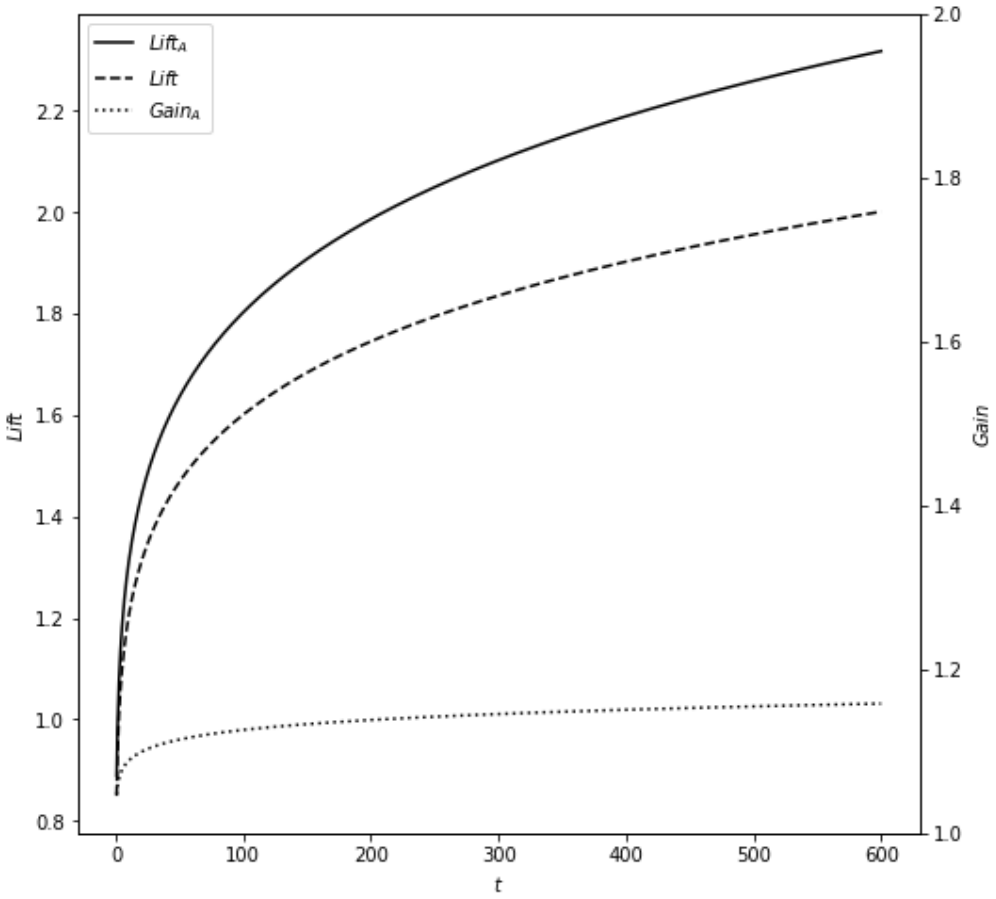

Figure 16. Lifts and gain in the case study (Cox PH model).

Fixing maintenance interval to, for example, 300 days the Cox PH model based scheduling would produce 13% less unplanned visits, than the current scheduling (gain being 1.15 at 300 days). To see this, recall that lift yields the ratio of unplanned visit rates of an average elevator and a freshly maintained elevator. Now, let us denote that a maintenance visit reduces the rate of future unplanned visits by factor X in all elevators group and factor X_A in the high rank group. Writing $X = 1 / Lift$ and $X_A = 1 / Lift_A$, yields $Gain = X / X_A$. Gain being 1.15 at 300 days means that the ratio of expected rates of unplanned visits right after maintenance in the two groups is $X_A / X = 1 / 1.15$ which is around 0.87. In other words, expected value of unplanned visits in the high rank group is 13% less than that in the all elevators group.

4.3.2 Lifts from WTTE-RNN model

Similar exercise was done with WTTE-RNN model. Figure 17 graphs the average of the survival curves for the high rank elevators (group A) and all elevators.

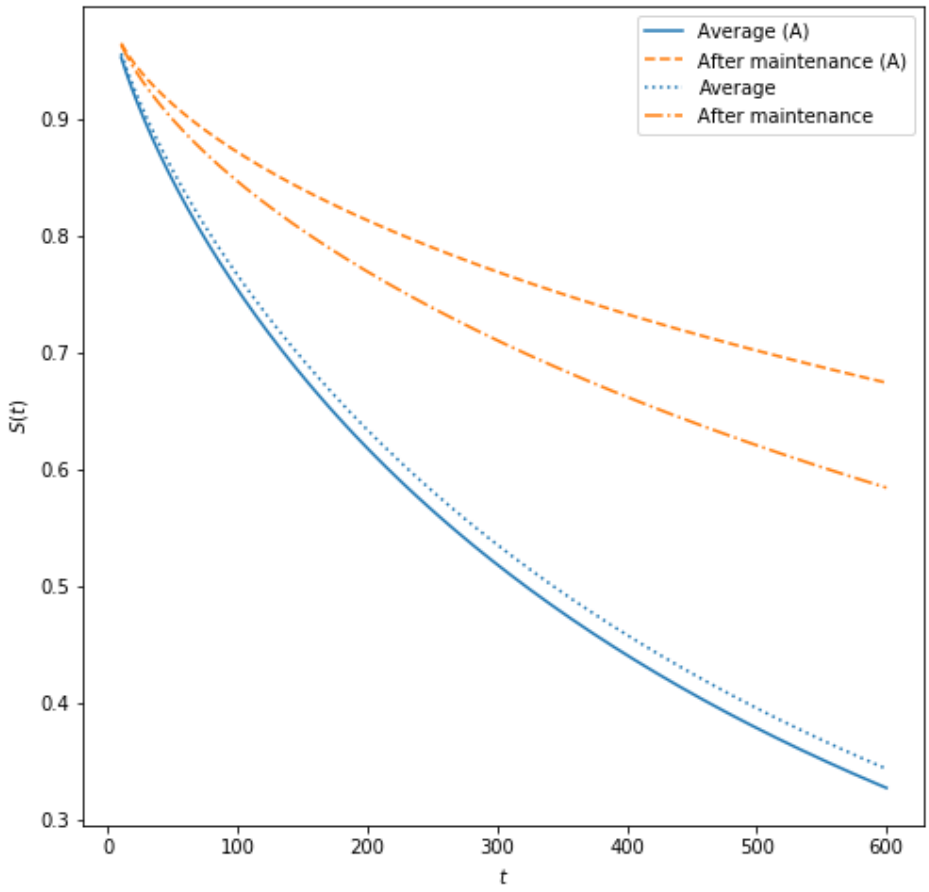

Figure 17. Survival curve estimates in the case study with WTTE-RNN model

Figure 18 shows corresponding lifts and gain for the WTTE-RNN model. The gain for 300 day maintenance interval is 1.39 as compared to that of 1.15 obtained with the Cox PH model. Respective reductions in the amount of unplanned visits are 28% and 13%.

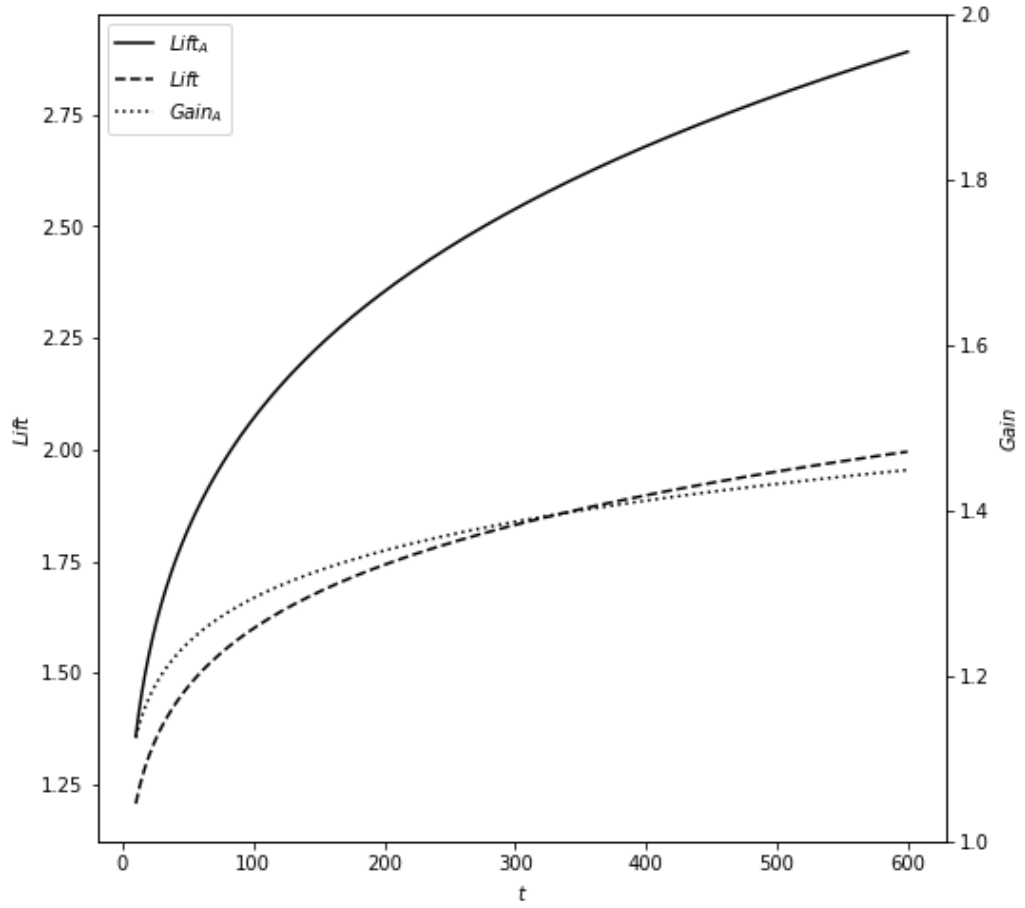

Figure 18. Lifts and gain in the case study (WTTE-RNN model).

5 CONCLUSIONS

The aim of the thesis was to study how elevator doors respond to maintenance and how a prognostic model could be used to decrease the rate of elevator door malfunctions by scheduling the preventive maintenance visits according to model recommendations. Idea being that a prognostic model could be used to filter out elevators that would benefit most of a preventive maintenance visit.

Survival analysis methods were used to analyse available data that consisted of maintenance records and condition monitoring data (20 thousand elevators for a period of two years). Condition monitoring data was aggregated on a daily level. First, stratified sur-

vival curves were obtained by fitting a non-parametric Kaplan-Meier estimator with the maintenance records. This was done for two groups: all elevators and freshly maintained elevators. In addition, the so-called Weibull plot technique was used to establish that survival times roughly adhere to the Weibull distribution. This justified performing stratified Weibull fit for the two groups, which gave a nice continuous visualization of how a maintenance visit generally improves door malfunction rate (research question 1), and how the value of a preventive maintenance visit develops over time (research question 2).

Next, a real-life dataset was built using both the maintenance records and condition monitoring data and a semi-parametric model called Cox PH model was used to approximate the survival. As opposed to Kaplan-Meier estimator, a trained Cox model can be inferenced for survival using condition monitoring data. Performance of the model was evaluated using cross-validation and a metric called C-index. A comparative analysis to a more complex neural network model called WTTE-RNN was performed. It was established that the C-index for the models in saturation were 0.65 and 0.71 for the Cox PH model and the WTTE-RNN model respectively. This analysis answered the research question 3 (*"How well certain mathematical models can predict the risk of malfunction based on condition monitoring data?"*).

Finally, a method was put forward for decreasing the rate of door malfunctions via condition monitoring based scheduling of preventive maintenance. Method relies on first inferencing a prognostic model for daily survival estimates of all elevators since start of their current observation periods. These predictions are then used for evaluating drift in elevator condition. Elevators having the highest drift are thought to benefit most of a preventive maintenance visit. Research hypothesis was that if scheduling of preventive maintenance visits would be based on the model recommendations (i.e. the drift), rather than calendar based scheduling, then the rate of door malfunctions would be decreased. A case study was made to validate the research hypothesis using the proposed method with historical data. Both the Cox PH model and the WTTE-RNN were evaluated.

In the case study, prognostic model recommendations were used for selecting daily high rank elevators. Then, stratified survival curves were calculated by performing a Weibull

fit for the freshly maintained elevators and all elevators. This was done separately for the high rank elevators and all elevators (i.e. four survival curves were obtained altogether). A function called *Lift* was defined to capture the maintenance effect as a function of time (i.e. how a preventive maintenance visit generally enhances survival within the particular group of elevators). Using the four survival curve estimates, two lift curves were formed for high rank elevators, and all elevators respectively. Lastly, a concept called *Gain* was used for comparing the maintenance effect within the high rank elevators and all elevators. A *Gain* curve was obtained by division of the two lift curves. From the *Gain* curve, a result was derived, that the proposed scheduling method would decrease the door malfunction rate by 13% using the Cox PH model and by 28% using the WTTE-RNN model (when maintenance interval of 300 days was assumed).

Since the case study was made in retrospect, it is not possible to anticipate the exact change in the rate of reported door malfunctions should the proposed method be used in real-life. In the beginning, the algorithm could be able to select elevators whose doors would benefit significantly from maintenance. However, it is not clear how soon such high lift elevators would run out, leaving scheduling algorithm with a fairly homogenous group of elevators to select from. Another practical point to consider is that the loss function of the WTTE-RNN model relies on censoring process being non-informative. Now, if scheduling decisions would be based on the WTTE-RNN model, then censoring would no longer be strictly non-informative (as preventive visit causes censoring). Observation periods that end due to preventive maintenance visit being scheduled by the model should in theory be dismissed from future training sets. This could limit the usefulness of the method. Due to these reasons, it would be subject for further work to quantify the real-life benefit of leveraging the methods discussed in the thesis. A comparison between the proposed condition monitoring based scheduling and calendar based scheduling would provide clearly quantified results on the efficacy and applicability of the proposed method.

REFERENCES

Armitage P., Berry G., Matthews J.N.S. 2002, *Statistical methods in medical research*, 4th ed., Blackwell Publishing, pp. 85-86.

Bland, J.M. & Altman, D.G. 2004, The logrank test: *British Medical Journal*, Vol. 328, Issue 7447, p. 1073.

Bousdekis, A. et al. 2018, Review, analysis and synthesis of prognostic-based decision support methods for condition based maintenance: *Journal of Intelligent Manufacturing*, Vol. 29, No. 6, pp. 1303–1316.

Castro, I. T. et al. 2012, A predictive maintenance strategy based on mean residual life for systems subject to competing failures due to degradation and shocks: *11th International Probabilistic Safety Assessment and Management Conference and the Annual European Safety and Reliability Conference 2012*, PSAM11 ESREL 2012, 1, pp. 375–384.

Christ, M. & Braun, N. & Neuffer, J. & Kempa-Liehr A.W. 2018, Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh -- A Python package): *Neurocomputing 307*, pp. 72-77.

Cox, D.R. 1972, Regression Models and Life-Tables: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 34, No. 2, pp. 187–220.

Cox, D.R. 1975, Partial likelihood: *Biometrika*, Vol.62, No. 2, pp. 269–276.

De Laurentiis, M., Ravdin, P.M., (1994), 'Survival analysis of censored data: Neural network analysis detection of complex interactions between variables.', Breast Cancer Research Treatment 32, pp. 113–118.

Deep-TTF 2018. Available from https://github.com/gm-spacagna/deep-ttf/, Accessed 3.3.2020.

Dittman, D. and Khoshgoftaar, T. (2014) 'Comparison of Data Sampling Approaches for Imbalanced Bioinformatics Data', The Twenty-Seventh …, pp. 268–271. Available at: http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS14/paper/viewFile/7850/7850

Giunchiglia, E. & Nemchenko, A. & van der Schaar, M. 2018, RNN-SURV: A deep recurrent model for survival analysis: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11141 LNCS, pp. 23–32.

Harrell, F.E. et al 1982, Evaluating the yield of medical tests: *JAMA - Journal of the American Medical Association*, Vol. 247, No. 18, pp. 2543–2546.

Jing, B. et al. 2019, A deep survival analysis method based on ranking: *Artificial Intelligence in Medicine, Elsevier*, Vol. 98, pp. 1–9.

Kaplan, E. L. and Meier, P. (1958) 'Nonparametric estimation from incomplete samples', J. of the ASA, 73(282), pp. 457–481. Available at: https://web.stanford.edu/~lutian/coursepdf/KMpaper.pdf.

Lee, H. et al. (2010) 'Toward optimal churn management: A partial least square (PLS) model', 16th Americas Conference on Information Systems 2010, AMCIS 2010, 2(July), pp. 961–971. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84870447433&partnerID=40&md5=de6940dc252b840c89af555ff5d049be.

Leontjeva, A. & Kuzovkin, I. 2016, Combining static and dynamic features for multivariate sequence classification: *Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics*, DSAA 2016, pp. 21–30.

Lifelines. 2019. Available from https://lifelines.readthedocs.io/en/latest/ Accessed 3.3.2020.

Mantel, Nathan. 1966. 'Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration.' Cancer Chemotherapy Reports 50 (3): 163–70.

Martinsson, E. (2016) 'WTTE-RNN : Weibull Time To Event Recurrent Neural Network A model for sequential prediction of time-to-event in the case'. Available at: https://ragulpr.github.io/assets/draft_master_thesis_martinsson_egil_wtte_rnn_2016.pdf

Nalchigar, S. & Yu, E. 2018, Business-driven data analytics: A conceptual modeling framework: *Data & Knowledge Engineering*, Elsevier Ltd, Vol. 117, pp. 359–372.

Nelson, W.B. 2004, *Applied Life Data Analysis*, John Wiley & Sons, 664 pages.

Ohno-Machado, L. 1996, Sequential use of neural networks for survival prediction in AIDS: *Proceedings of a conference of the American Medical Informatics Association,* AMIA Fall Symposium, pp. 170–174.

Raykar, V.C. et al. 2008, On ranking in survival analysis: Bounds on the concordance index: *20th NIPS*, pp. 1209–1216.

Schober, P. & Vetter, T.R. 2018, Survival analysis and interpretation of time-to-event data: The tortoise and the hare: *Anesthesia and Analgesia*, Vol. 127, No. 3, pp. 792–798.

Scholz, F. (2008) 'Inference for the Weibull Distribution', Stat 498B Industrial Statistics, 632, p. 59. Available at: http://www.stat.washington.edu/fritz/DATAFILES498B2008/WeibullBounds.pdf.

Scikit-survival. 2019. Available from https://github.com/sebp/scikit-survival/ Accessed 3.3.2020.

Vanderplas et al (2012), 'Introduction to astroML: Machine learning for astrophysics', proc. of CIDU, pp. 47-54.

Weibull, W. 1951, A statistical distribution function of wide applicability: *ASME Journal of Applied Mechanics*, Vol. 18, No. 3, pp. 293–297.

Widrow, B. 2005, Thinking about thinking: The discovery of the LMS algorithm: *IEEE Signal Processing Magazine*, Vol. 22, No. 1, pp. 100–103.

WTTE-RNN 2019. Available from https://github.com/ragulpr/wtte-rnn/ Accessed 3.3.2020.

Wolstenholme, L.C. 1999, *Reliability Modelling: A Statistical Approach*, Chapman and Hall/CRC, 256 pages.

Wu, S. J. et al. 2007, A neural network integrated decision support system for condition-based optimal predictive maintenance policy: *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, Vol. 37, No. 2, pp. 226–236.