Duale Hochschule Baden-Württemberg Mannheim

Seminararbeit

Data Exploration - Projektreport Survival Analysis

Studiengang Wirtschaftsinformatik

Studienrichtung Data Science

Verfasser(in):	Lifesaver (Lennart Schulz, Marcel Winter, Laura Struss)
Matrikelnummer:	3300490, 5542090, 4212678
Kurs:	WWI19DSA
Bearbeitungszeitraum:	11.05.2021 – 13.07.2021

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

BMI	Body Mass Index
UCLA	University of California, Los Angeles

1 Theoretische Grundlagen

1.1 Thema und Motivation

Das menschliche Leben wird häufig durch Krankheiten beeinflusst oder gar beendet. In diesem Projekt geht es um die Zusammenhänge zwischen einer Krankheit und bestimmten Attributen, sodass Vorsorgemaßnahmen angepasst werden können. Dadurch wird sich eine Steigerung der Lebensqualität sowie ein optimierter Umgang mit Krankheiten und weiteren unumgänglichen Ereignissen erhofft. Es soll eine Prognose über den möglichen Zeitraum des Todes einer Person mit einer bestimmten Krankheit getroffen werden, sodass diese Person ihre restliche Lebenszeit mit mehr Gewissheit verbringen kann. Die hauptsächliche Motivation liegt darin, durch die erkannten Zusammenhänge zwischen Attributen und bestimmten Krankheiten rechtzeitig Maßnahmen einleiten zu können, sodass ein schwerwiegender oder gar tödlicher Verlauf einer Krankheit verhindert werden kann. Um diese Motivation zu verkörpern, nannte sich das Projektteam "Lifesaver", denn dieses Projekt soll Leben retten.

1.2 Related Work

Das Projekt basiert auf der Studie “Applied Survival Analysis: Regression Modeling of Time-to-Event Data“. In dieser Studie wurden 500 Patienten dokumentiert, die nach einem Herzinfarkt ins Krankenhaus eingewiesen wurden. Die Anamnese wurde in die Studie mit aufgenommen, um den Krankheitsverlauf bis zu Austritt aus der Studie zu dokumentieren und bewerten zu können (vgl. David W. Hosmer Jr., 2008, Z.1 ff.).

2 Praktische Grundlagen

2.1 Verwendete Technologien und Bibliotheken

Das Projekt wurde mittels verschiedener Bibliotheken und technologien erstellt, welche im Folgenden aufgeführt und erläutert werden. Die Numpy-Bibliothek wurde für simple bis komplexe mathematische Operationen verwendet. Für die Daten-Analyse und -Manipulation wurde die Pandas-Bibliothek verwendet. Zur statistischen Datenvisualisierung wurden die Bibliotheken Matplotlib und Seaborn verwendet. Für den Umgang mit Warnungen wurde die interne Python-Bibliothek Warnings genutzt. Die umfangreiche Bibliothek Lifelines wurde für die Survival Analyse benutzt und mit einfachen built-in Funktionen von Matplotlib und weiteren kombiniert.

2.2 Präsentation der Ergebnisse

Nachdem alle benötigten Bibliotheken importiert wurden, wurden die Daten transformiert. Dafür wurden die Werte nahezu aller Spalten zunächst in den Datentyp Integer umgewandelt. Anhand der Daten wurde anschließend ein erster Test vorgenommen mittels des Kaplan-Meier-Modells.

Das Kaplan-Meier-Modell ist ein nicht-parametrisches univariates Modell zur Schätzung der Überlebensfunktion. Das Modell ist eines der simpelsten Verfahren zur Ereignis-Zeit-Analyse und wurde ursprünglich für die medizinische Statistik, wie in diesem Projekt angewendet, entwickelt. Es bezieht sich hierbei lediglich auf die Teilnahmedauer der Studie sowie auf das Eintreten des zu untersuchenden Ereignisses, wie in diesem Fall den Tod. Aufgrund der Einfachheit des Verfahrens wurde das Modell bei dem ersten Test verwendet. Sämtliche Survival Analysis Modelle besitzen stets eine Überlebensfunktion, eine kumulierte Risikofunktion und eine Risikofunktion (vgl. Davidson-Pilon et al., 2021, Z.1 ff.).

Nach einem ersten Test wurden die Daten deskriptiv analysiert. Hierzu wurde der Durchschnitt, die Standardabweichung, die Varianz und die Anzahl der einzigartigen Werte sämtlicher Attribute untersucht. Außerdem wurde eine Korrelationsmatrix erstellt, aus welcher Attributspaare mit hoher positiver oder negativer Korrelation herausgefiltert wurden. Nach einigen Tests mittels des Cox Probability Hazard Modells, welches im späteren Verlauf noch erklärt wird, konnten die fünf signifikantesten Attribute herausgefiltert werden.

Um das beste Survival Analysis Modell herauszufinden, wurde ein bei Lifelines bereits integrierter grafischer sowie mathematischer Test für univariate parametrische Modelle

durchgeführt. Daraus ergibt sich, dass kein univariates parametrisches Modell für diese Daten ausreichend gut geeignet ist (vgl. Davidson-Pilon et al., 2021, Z.1 ff.).

Neben univariaten Modellen gibt es sogenannte Survival Regression Modelle, welche die Auswirkungen verschiedener Attribute auf den Krankheitsverlauf analysieren. Eines der bekanntesten Modelle dafür ist das Cox PH Modell, welches semi-parametrisch ist und somit die Eigenschaften von parametrischen und nicht-parametrischen Modellen kombiniert. Lifelines bietet umfangreiche Möglichkeiten zur Anpassung des Modells, wie beispielsweise das Einstellen eines Penalizers oder Stratifikation. Außerdem gibt es in Lifelines diverse Methoden zur Validierung der Güte des Modells, die sowohl durch Lifelines automatisch im Hintergrund durchgeführt werden, als auch manuell angewendet werden können (vgl. Davidson-Pilon et al., 2021, Z.1 ff.).

Die Güte der Vorhersage kann durch verschiedene Messwerte bestimmt werden, diese werden im Folgenden aufgeführt und erläutert.

Der Concordance Index ist eine erweiterte Form der Area under the Curve (AUC). Dieser Index bewertet die Genauigkeit der Rangfolge der vorhergesagten Zeit. Ein Wert von 0 bedeutet, dass das Modell die Daten optimal invers vorhersagt. Ein Wert von 0,5 bedeutet, dass das Modell zufällig bewertet und ein Wert von 1 bedeutet eine optimale Vorhersage (Overfitting). Die Ergebnisse des angewendeten Modells liegen bei 0,77 und liegen somit im Normalbereich (vgl. Davidson-Pilon et al., 2021, Z.1 ff.).

Eines der am häufigsten benutzten Kriterien zur Auswahl eines Modells in der Survival Analysis ist das Akaike-Informationskriterium (AIC). Dies gibt außerdem die relative Qualität von statistischen Modellen von einem gegebenen Datensatz an. Das Modell hat schlussendlich einen partiellen AIC von 2248,75 erreicht, dies ist der niedrigste Wert, den das Modell nach ausgiebigen Iterationen erzielt hat (vgl. Davidson-Pilon et al., 2021, Z.1 ff.).

Lifelines bietet eine Cross-Valididierungsfunktion, welche das Aufsplitten der Trainingsdaten überflüssig macht. Diese Funktion wurde in diesem Projekt angewendet und ergab einen Concordance Index von 0,8. Dieser Wert liegt an der Grenze zum Overfitting.

Bei einer Sterberate von 43 Prozent wurden mithilfe der Survival Probability Calibration vorhergesagte und beobachtete Wahrscheinlichkeiten ermittelt. Diese Wahrscheinlichkeiten sind in Abbildung 2.1 veranschaulicht. Je näher die rote Kurve an der schwarzen Kurve liegt, desto besser ist das Modell. Wie in der Abbildung 2.1 zu sehen ist, liegt die rote Kurve sehr nah an der schwarzen Kurve, sodass das Modell nah am Optimum liegt.

Die Überlebensfunktion des Cox PH Modells entspricht den Erwartungen und bildet die Überlebenswahrscheinlichkeit in Abhängigkeit von der Zeit in Prozent ab. Die dazugehörige kumulierte Risikofunktion gibt dementsprechend das kumulierte Risiko für die entsprechenden Tage an. Diese Funktion gibt die Wahrscheinlichkeit dafür an, bis zu einem Tag X

[1] Eigene Darstellung

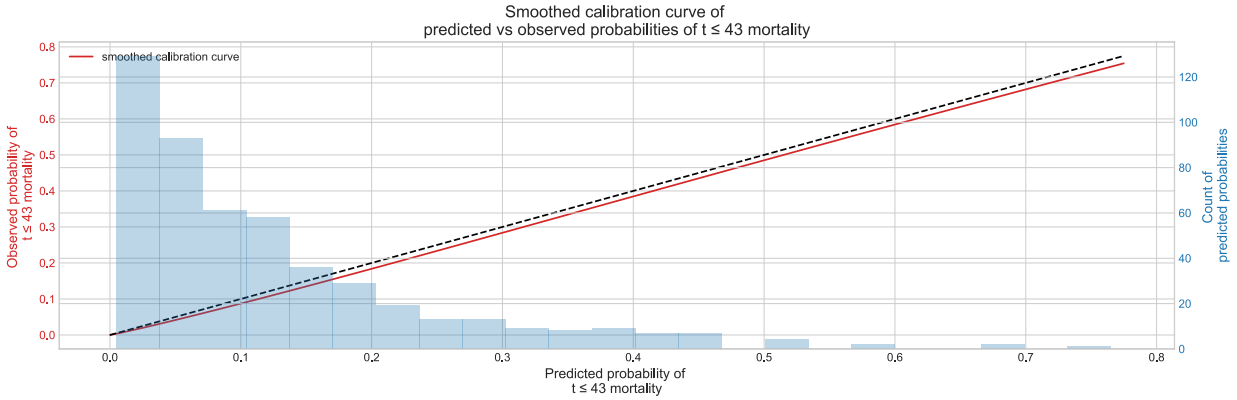

Abbildung 2.1: Survival Probability Calibration[1]

gestorben zu sein. Die Risikofunktion, welche die Ableitung der kumulierten Risikofunktion ist, gibt an, wie hoch die Wahrscheinlichkeit ist, an diesem Tag zu sterben.

In der Abbildung 2.2 sieht man den Einfluss des Geschlechts auf die Überlebenswahrscheinlichkeit, basierend auf dem gewählten Datensatz. Man erkennt deutlich, dass Frauen eine prozentual höhere Wahrscheinlichkeit dafür haben, eine Herzkrankheit zu überleben.

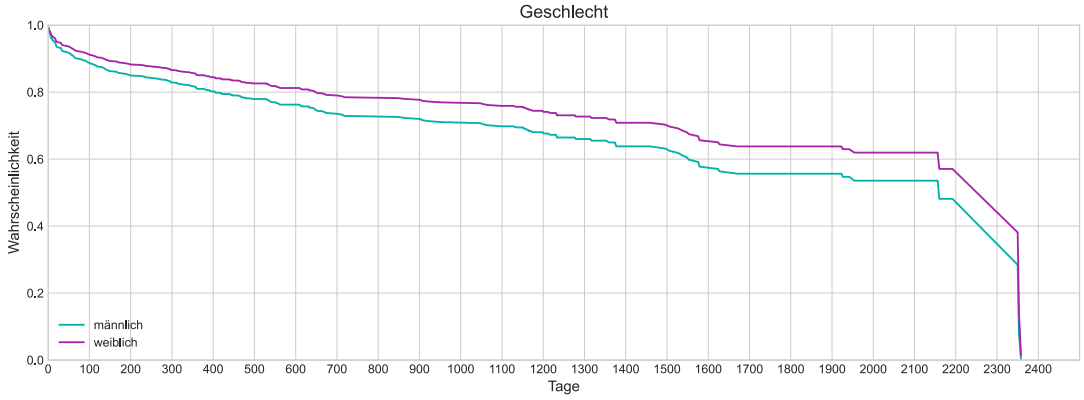

Abbildung 2.2: Impact of the gender on the Survival Probability [2]

[2]Eigene Darstellung

2.3 Kritische Bewertung der Ergebnisse

Aufgrund der höheren Komplexität der Survival Analysis im Gegensatz zur normalen Regression, ergeben sich mehrere Faktoren, welche beachtet werden müssen. Dazu zählt unter Anderem die Beachtung nicht-linearer Korrelationen der Attribute mit der Überlebenswahrscheinlichkeit. Darunter versteht man beispielsweise, wie in Abbildung 2.3 zu sehen ist, dass der BMI laut Modell eine lineare Korrelation zur Überlebenswahrscheinlichkeit hat. Das Modell geht davon aus, dass je höher der BMI desto höher die Überlebenswahrscheinlichkeit ist. Diese Annahme ist faktisch falsch, da bei einem sehr hohen BMI häufig ein erhöhtes Sterblichkeitsrisiko vorliegt (vgl. Hauner, 2009, Z.1 ff.). Daraus ergibt sich, dass das Modell hier keine richtigen Vorhersagen trifft. Es müssten weitere Korrelationsanalysen für nicht-lineare Zusammenhänge gemacht werden sowie die Regressionsformel für dieses Attribut angepasst werden, damit das Modell damit korrekt umgehen kann. Da dies erst zu einem späteren Zeitpunkt aufgefallen ist und die Umsetzung dieser Korrelationsanalysen den Umfang dieser Arbeit sprengen würde, wurde das Attribut nicht weiter berücksichtigt (vgl. Davidson-Pilon et al., 2021, Z.1 ff.).

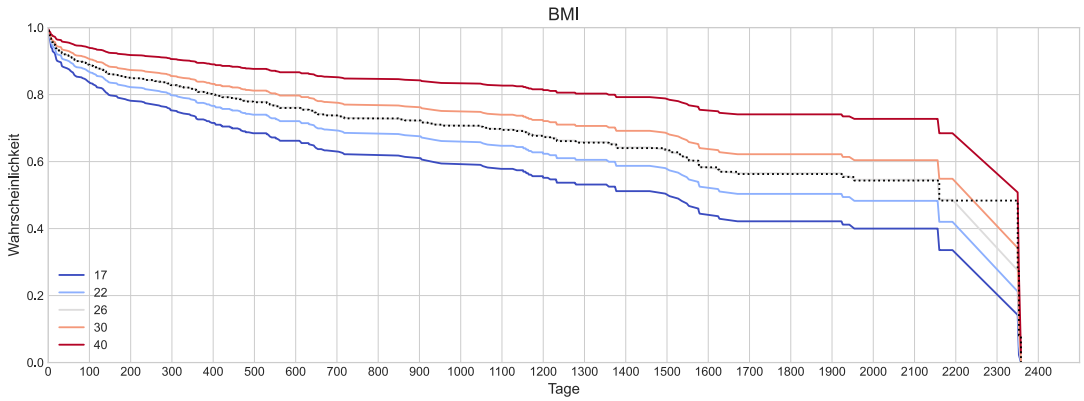

Abbildung 2.3: Impact of the gender on the Survival Probability [3]

[3] Eigene Darstellung

Survival Analysis Modelle gehen von einer abfallenden Wahrscheinlichkeit aus, was bedeutet, dass die Kurve stets monoton fallend ist. Außerdem nimmt das Konfidenzintervall bei zunehmender Zensierung der Daten zu, was die Vorhersagegenauigkeit stark einschränken kann und, wie in diesem Beispiel zu sehen ist, ab 2200 Tagen zu einem starken Abfall der Überlebenswahrscheinlichkeit führen kann. Daraus resultiert, dass die Interpretation der Ergebnisse ungenauer werden, je weiter man sich rechts auf der X-Achse befindet (vgl. Davidson-Pilon et al., 2021, Z.1 ff.).

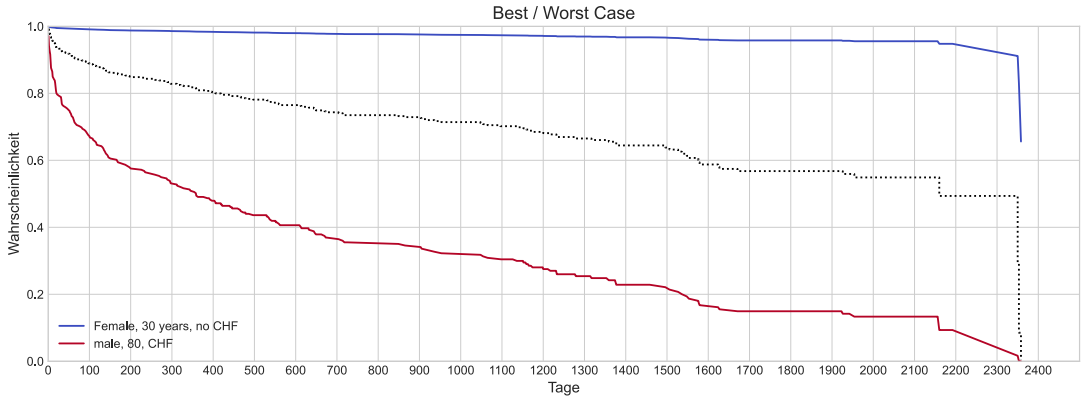

Abbildung 2.4: Survival Probability difference between the best and the worst case[4]

Survival Analysis ist ein komplexes Thema, welches weiter geht als normale Regression, da bspw. Dinge wie Zensierungen beachtet werden müssen. Bei der Auswahl der Modelle gibt es einige Dinge zu beachten. Zum Einstieg und zur Voranalyse der Daten bietet es sich an, univariate Modelle wie das Kaplan-Meier-Modell zu nutzen. Möchte man allerdings Auswirkungen von Attributen auf den Krankheitsverlauf berücksichtigen, sollte man Modelle wie das Cox-PH-Modell nutzen. In der Lifelines-Bibliothek gibt es verschiedene Möglichkeiten zur Konfiguration der Modellberechnungen, wie Penalizer, Stratification und weitere statistische und machine-learning-spezifische Faktoren. Außerdem gibt es verschiedene Methoden zur Validierung der Modelle sowie grafische Möglichkeiten zur Darstellung der Auswirkungen von Attributen auf den Krankheitsverlauf. In diesem Projekt wurde bei der Analyse von 500 Patienten nach Einweisung ins Krankenhaus aufgrund eines Herzinfarkts schlussendlich festgestellt, dass das Geschlecht, Alter, sowie das Vorhandensein einer Herzinsuffizienz, der Blutdruck, als auch die Herzfrequenz eine signifikante Rolle beim Krankheitsverlauf spielen und die Überlebenswahrscheinlichkeit stark verbessern oder verschlechtern können.

Aufgrund dieser gefundenen Zusammenhänge zwischen verschiedenen Attributen und einer Krankheit können Krankenhäuser in Zukunft Risikopatienten besser erkennen und sich besser auf diese fokussieren. Die bereits in der Medizin bekannten Risikofaktoren, wie beispielsweise hohes Alter, konnten innerhalb dieses Projektes belegt werden.

[4] Eigene Darstellung

A Anhang

Name	Beschreibung	Anmerkung
AGE	Alter zum Zeitpunkt der Krankenhauseinweisung	
GENDER	Geschlecht (m oder w)	
HR	Herzfrequenz in Schlägen pro Minute	
SYSBP	Systolischer Blutdruck	oberer Messwert, während der Anspannungs- und Auswurfphase der linken Herzkammer maximal entwickelter Druck Normbereich: 110-130 mmHg
DIASBP	Diastolischer Blutdruck	unterer Messwert, niedrigster Wert während der Entspannungs- und Erweiterungsphase des Herzmuskels; Phase zwischen größter Druckentwicklung (systolischer Druck) und größtem Druckabfall (diastolischer Druck) wird als Diastole bezeichnet.
BMI	Body Mass Index	
CVD	Kardiovaskuläre Erkrankung	
AFB	Vorhofflimmern	Herzrhythmusstörung
SHO	Kardiogener Schock	kann zu Multiorganversagen und zum Tod führen
CHF	Herzinsuffizienz	
AV3	Av-Block 3. Grades	
MIORD	Herzinfarkt	0= First, 1= Wiederkehrend
MITYPE	Herzinfarkt Typ	0 = Nicht-Wellen Herzinfarkt 1 = Wellen Herzinfarkt
YEAR	Jahr der Einweisung	1997, 1999, 2001
LOS	Länge des Krankenhausaufenthalt	in Tagen
LENFOL	Teilnahmedauer in Tagen bis Eintreten des Events bzw des Austritts des Teilnehmers	
FSTAT	Ist das Event eingetreten? Event = Tod	

Literaturverzeichnis

David W. Hosmer Jr., e. a. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Wiley. https://www.wiley.com/en-us/Applied+Survival+Analysis%C3%A7%3A+Regression+Modeling+of+Time+to+Event+Data%5C%2C+2nd+Edition-p-9780471754992

Davidson-Pilon, C., Kalderstam, J., Jacobson, N., Reed, S., Kuhn, B., Zivich, P., Williamson, M., AbdealiJK, Deepyaman Datta, Fiore-Gartland, A., Parij, A., WIlson, D., Gabriel, Moneda, L., Moncada-Torres, A., Stark, K., Gadgil, H., Jona, JoseLlanes, ... Skipper Seabold. (2021). *CamDavidsonPilon/lifelines: 0.26.0*. Zenodo. https://doi.org/10.5281/zenodo.4816284

Hauner, Ä. H. (2009). *Übergewicht - alles halb so schlimm?* Verfügbar 12. Juli 2021 unter https://www.aerzteblatt.de/archiv/66140/Uebergewicht