Stockholm University Demography Unit,
Dissertation series 34

Stockholm University

Life Course Prediction for Labor Market Outcomes

Lessons from Register-Based Machine Learning

Sanni Kuikka

Doctoral Thesis in Sociological Demography at Stockholm University, Sweden 2026

Life Course Prediction for Labor Market Outcomes

Lessons from Register-Based Machine Learning

Sanni Kuikka

Academic dissertation for the Degree of Doctor of Philosophy in Sociological Demography at Stockholm University to be publicly defended on Thursday 19 November 2026 at 13.00 in hörsal 5, hus B, Universitetsvägen 10 B and online via Zoom, public link is available at the department website.

Abstract

Machine learning has begun making its way into the methodological toolkit for the social sciences, but only very recently, with many questions still unanswered. Combining this toolkit with population register data offers a promising perspective into researching the human life course and the complexities within. This dissertation explores how Swedish and Finnish population registers combined with decision tree-based machine learning algorithms can be leveraged for labor market research, in particular for studying the predictability of long-term unemployment in early adulthood and lifetime disposable income. The theoretical foundations on this work lie at the intersection of the life course framework and how different forms of capital accumulate and transfer intergenerationally. Machine learning offers a useful tool to jointly model multiple domains in life, and how they together form trajectories towards these labor market outcomes.

Studies 1 and 2 concern long-term unemployment in early adulthood in the context of Finland and leverage data from the full birth cohort of 1987. The predictors for both studies are chosen based on theories and previous empirical literature on unemployment, and operationalized longitudinally where possible, some following the ages related to school transitions. Study 1 focuses on the extent to which long-term unemployment at ages 25–30 can be predicted, as well as the differences in predictability across the outcome groups and socioeconomic background. The results show that when compared to logistic regression, all the tree-based models fare better overall, but their performance is mostly due to correctly predicting the outcome class employed, while incorrectly predicting most of the unemployed. Regarding the socioeconomic origin, the low socioeconomic status (SES) group is more difficult to predict overall than the high-SES group, but within the low-SES group, the correct predictions for the unemployed increase. The findings highlight our limited understanding of the processes leading to unemployment, as well as the importance of learning more about sub-group predictability. Study 2 unpacks these results for the best performing model, Gradient Booster, by exploring the combinations that predict someone as: A) employed or unemployed, and B) correctly or incorrectly employed or unemployed. The aim is to further evaluate which processes and theories are the model's blind spots, and the method used is Principal Component Analysis. For the employed group, parental unemployment was the prominent component steering the model wrong, while for the unemployed it was the combination of family cohesion, mental wellbeing and intergenerational transmission of education.

Study 3 focuses on lifetime disposable income in the context of Sweden, leveraging population registers for the birth cohorts of 1967–1971. This study explores the extent to which it is possible to predict a person's income class based on their early life conditions and further unpacks this to answer which characteristics each income class shares, and which sub-groups end up in the same income class, through different combinations of characteristics. The predictors are based on theories and empirical literature, not only focusing on socioeconomic origins, but also geographic origins, family structure, and gender. The analysis of predictability compares logistic regression to two tree-based algorithms, as well as a composite model where a different model for each income group is chosen based on highest performance. The results find the machine learning models to outperform multinomial logistic regression, with the composite model being the best. The middle-income class has the highest predictive accuracy, while the lowest income class has the lowest. These predictions are then unpacked using Shapley Additive Explanations, which are further analyzed using Hierarchical Agglomerative Clustering. The results reveal heterogeneity both across and within income classes, with complex realities at the bottom of the income distribution, and a particularly streamlined path at the top.

This dissertation contributes to the field of life course prediction by exploring the predictability of the two labor market outcomes and highlights the importance of exploring the predictability of sub-groups. Furthermore, it presents ways in which population register data can be analyzed with machine learning methods to reveal more details about the human life course.

Keywords: *life course, machine learning, labor market, unemployment, income, population registers, register-based, life course prediction.*

Stockholm 2026
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-260353

ISBN 978-91-8107-753-7
ISBN 978-91-8107-754-4
ISSN 1404-2304

Department of Sociology

Stockholm University, 106 91 Stockholm

Stockholm University

LIFE COURSE PREDICTION FOR LABOR MARKET OUTCOMES

Sanni Kuikka

Stockholm University

Life Course Prediction for Labor Market Outcomes

Lessons from Register-Based Machine Learning

Sanni Kuikka

©Sanni Kuikka, Stockholm University 2026

ISBN print 978-91-8107-753-7
ISBN PDF 978-91-8107-754-4
ISSN 1404-2304

Cover image: “Arbres en Île-de-France” by Armand Guillaumin (circa 1895).
Sourced via Artvee / Private Collection. Public Domain.

Printed in Sweden by Universitetsservice US-AB, Stockholm 2026

I know you have
a little life in you yet
I know you have
a lot of strength left

This Woman's Work,
Kate Bush, 1989

Contents

Acknowledgements 1

Sammanfattning 4

List of studies 6

Introduction 7
- Machine learning, a new era of social inquiry 7
- Life course prediction with population register data 9
- Aims 12

Theoretical and empirical context 14
- Life course framework 14
- Different forms of capital 16
- Long-term unemployment in early adulthood 16
- Lifetime disposable income 17

Methods and data 20
- On statistical modelling cultures and sociological explanation 20
- Machine learning 21
  - Supervised machine learning 22
  - Unsupervised machine learning 24
  - Summary of the models 25
- Shapley values 26
- General remarks on life course prediction 27
- Population register data 28
  - Finland 29
  - Sweden 30
  - On research infrastructure 31

Ethical considerations 34

Summary of studies 36
- Study 1 - The (Un)Predictability of Early (Un)Employment: A Machine Learning Approach 36
- Study 2 - Unpacking the Unpredictability of Early Unemployment: Exposing the Blind Spots of a Machine Learning Model 37

Study 3 - Diverging Paths to Economic Destination: Uncovering Heterogeneity in Early Life Conditions Across and Within Income Classes 37

Discussion and conclusion 39

References 41

Acknowledgements

I never thought I would end up doing a PhD: I always thought it was for other people, those intellectual, tenacious, and determined ones. Somehow, I ended up here regardless. To get to this point, it has not only taken a lot of work, but also lots of support and encouragement from many people, to whom I now want to express my gratitude.

My main supervisor, Maria Brandén: you have been involved in my path ever since 2018 when I started my master's degree at IAS, and you were assigned as my mentor based on our similar research interests. You've been supporting my progress ever since, by being my master's thesis advisor, by encouraging me to apply for a PhD position, and by constantly motivating me to keep exploring this uncharted line of research. My co-supervisor, Stefanie Möllborn: you joined the team some years later but quickly became an integral part of my progress. With your ability to zoom in and out between details and larger contributions in my work, you have truly helped me navigate to the goal. My second co-supervisor, Max Thaning: we first became co-authors and then somewhere along the line also friends (or was it the other way around?). Later, you joined my supervisor team and kept challenging my thinking in the best possible way. I want to thank all of you for your commitment to my work, but also your support regarding my wellbeing during this process. In addition to the supervisors above, I have also received supervision from more people at different stages of the process: for this I want to thank Siddartha Aradhya, Kieron Barclay, and Adel Daoud.

When I first moved to Stockholm where I had minimal social networks, there were two individuals at the department who took me in and made me feel particularly welcome. They have also kept checking in on me throughout the years, making me feel like I belong: Sunnee Billingsley and Anna Lund, thank you. To Hernan Mondani, thank you for providing great feedback on my work as my final seminar opponent, but also for all the talks and emails regarding sloths, manatees, and more. The department as a whole has been a safe home base to pursue my research from, and for that a warm thank you goes to all the researcher colleagues as well as the administrative and technical staff.

During my time as a PhD candidate, I have been surrounded by a large group of amazing peers whose PhD process has overlapped with mine: Amanda

Almstedt Valldor, Lovisa Backman, Abrar Bawati, Liliana Calderón, Erik Carlsson, Daniel Dahl, Pär Dalén, Jasper Darin, Itzel Díaz, Daniel Enström, Johanna Finnström, Elida Izani Ibrahim, Zeth Isaksson, Malcolm Jacobson, Kristin Jung Ståhle, Konstantin Kazenin (R.I.P.), Melis Kirgil, Tove Klasson, Weiwen Lai, Andreas Ljungström, Ida Ljungström, Mimmi Lounela, Nicole Rodriguez, Lea Schomaker, Ryan Switzer, Ane Tømmerås, Olivia Öhman Granström—the list goes on. Thank you all for sharing this journey with me: for the talks we've had, the conference trips, the hugs, the AWs, the vents. Flore Debruyne: you've been an amazing officemate, and I will miss the talks, hugs and tears we have shared over life and work. Signe Svallfors: you started as a great PhD buddy and continued as a great friend. Wooseong Kim: your smile and friendliness are some of the first things I remember when arriving at the department - sorry for the extra stress I caused by asking you to care for my plants. Frankseco Yorke: our venting sessions during the spring leading up to our final seminars were immensely supportive and they also led me to discover how similar we are as people in so many ways.

During my academic journey, which already started before the PhD project did, I have met people who have contributed to me pursuing this goal. Tiina Ristikari, thank you for hiring me as a summer intern at THL for my first exposure to the world of Nordic population register data research. Amber Beckley, thank you for hiring me as a research assistant in your project at SUDA, and encouraging me to apply for a PhD position. Matthew Salganik, thank you for your interest in my work, for our talks, and for inviting me to seminars. Anita Berglund, thank you for your passion in training young academics through the SINGS research school, and for your warmth towards me as a person behind my research. Maël Lecoursonnais, Enes Al Weswasi, and Miska Simanainen, thank you for shaping my SINGS experience to be a particularly friendly and fun one. Hailey Rheault-Huber, Ely Strömberg, and Emily Cantrell: thank you for the friendships that extend way beyond the academic life - it means a lot to me.

I would not have been able to keep my head together throughout this process by only being immersed in the academic world. Having friends outside the work I do has always been an important lifeline, and I am truly lucky to have so many people in my life to share things with. Veera, Taru, Santtu, Turo, Pete, Arttu, Kyra, Henrik H., Susanna, Christoffer, Evelina, Jacky, Lorenzo, Oscar, Gabbi, Alex: thank you all for being there for me through these years, either in person or online, during vacations or weekends, at music festivals, over boardgames, enjoying great food, in sickness and in health. I also want to thank my parents, Päivi and Esko. Bringing me up in the lack of economic capital, you still managed to pass on cultural capital in many forms: I believe it has been one of the carrying pillars throughout my life.

Moving to Sweden was never in my plans, but it has changed my life in ways I could not have imagined before. Sometimes I wonder if it was a coincidence at all that I ended up here, given that the women in my family have now migrated from Finland to Sweden in three generations: my grandmother as one of those some 80 000 war children during World War II, my mother as a labor migrant in the early 1980's, and me as a student in the late 2010's. The migrant experience was also passed down to me from my father's side, where both my grandparents were forced to evacuate Finnish Karelia as children, when the eastern border of Finland was re-drawn. The feeling of being out-of-place will likely always stay with me to some extent, but in Sweden I have also found a home and a family. Henrik, I cannot express in words how much you mean to me, and how much meeting you has changed my life for the better. You have never seen me without my role as a PhD candidate, without all the stress it has brought upon me, and yet you chose me as your partner. I am eagerly looking forward to co-parenting a tiny human with you, and I cannot imagine a better partner for that experience either. I am forever thankful and I love you so much.

Stockholm, October 2026

Sammanfattning

Maskininlärning har under senare år börjat bli en del av den samhällsvetenskapliga metodologiska verktygslådan, men många frågor är fortfarande obesvarade. Att kombinera maskininlärning med data från befolkningsregister har stor potential att avancera forskningen om människans livsförlopp och dess komplexitet. Denna avhandling undersöker hur svenska och finska befolkningsregister, i kombination med beslutsträdbaserade (*decision tree-based*) algoritmer för maskininlärning, kan användas för arbetsmarknadsforskning, specifikt för att studera förutsägbarheten av långtidsarbetslöshet i tidig vuxen ålder och disponibel inkomst under livet. De teoretiska grunderna för avhandlingen ligger i skärningspunkten mellan livsförloppsteori och hur olika former av kapital ackumuleras och överförs mellan generationer. Maskininlärning är ett särskilt användbart verktyg för att modellera flera livsområden simultant, och hur de i samspel med varandra bidrar till dessa arbetsmarknadsutfall.

Studie 1 och 2 handlar om långtidsarbetslöshet bland unga vuxna i Finland, och bygger på registerdata från hela kohorten födda 1987. De prediktorer som används i båda studierna har valts ut baserat på teorier och tidigare empirisk litteratur om arbetslöshet och har, där så är möjligt, operationaliserats longitudinellt. Studie 1 fokuserar på i vilken utsträckning långtidsarbetslöshet i åldern 25–30 år kan förutsägas, samt på skillnaderna i förutsägbarhet mellan olika utfallsgrupper och socioekonomiska bakgrunder. Resultaten visar att alla trädbaserade modeller överlag presterar bättre än logistisk regressionsanalys, men att deras goda prestation främst beror på att de är korrekta i sina förutsägelser för de som är sysselsatta, medan modellernas förutsägelser för långtidsarbetslösa ofta är felaktiga. När det gäller socioekonomisk bakgrund är gruppen med låg socioekonomisk status (SES) generellt sett svårare att förutsäga än gruppen med hög SES, men inom gruppen med låg SES ökar andelen korrekta förutsägelser för långtidsarbetslösa. Resultaten belyser vår begränsade förståelse för de processer som leder till arbetslöshet, liksom vikten av att lära oss mer om förutsägbarheten hos olika undergrupper. Studie 2 fördjupar analyserna för den modell som presterade bäst, Gradient Booster, genom att undersöka de kombinationer som förutsäger om en person är: A) sysselsatt eller långtidsarbetslös, och B) korrekt eller felaktigt klassificerad som sysselsatt eller långtidsarbetslös. Syftet är att ytterligare utvärdera vilka processer och teorier som utgör modellens blinda fläckar, och den metod som

används är principalkomponentanalys (*principal component analysis)*. För den sysselsatta gruppen var föräldrarnas arbetslöshet den huvudsakliga faktor som ledde modellen fel, medan det för de långtidsarbetslösa var familjesammansättning, psykisk hälsa och intergenerationell överföring av utbildning.

Studie 3 fokuserar på den disponibla livsinkomsten i svensk kontext och utgår från befolkningsregister för födelsekohorterna 1967–1971. Studien undersöker i vilken utsträckning det är möjligt att förutsäga en persons inkomstgrupp utifrån förhållandena under dennes tidiga liv, och fördjupar sig i vilka egenskaper som är gemensamma respektive skiljer sig åt inom respektive inkomstgrupp. Prediktorerna baseras på teorier och empirisk litteratur och fokuserar inte bara på socioekonomisk bakgrund, utan även på geografiskt ursprung, familjestruktur och kön. Analyserna av förutsägbarheten jämför multinomial logistisk regression med två trädbaserade algoritmer, samt en sammansatt modell där en separat modell väljs för varje inkomstgrupp utifrån bästa prestanda. Resultaten visar att maskininlärningsmodellerna presterar bättre än multinomial logistisk regression, där den sammansatta modellen är den bästa. Medelinkomstgruppen har den högsta prediktiva korrektheten (*accuracy*), medan den lägsta inkomstgruppen har den lägsta. Dessa förutsägelser bryts sedan ned med hjälp av Shapley Additive Explanations, och analyseras ytterligare med hjälp av hierarkisk agglomerativ klusteranalys. Resultaten visar heterogenitet både mellan och inom inkomstgrupperna, med komplexa förhållanden i den nedre delen av inkomstfördelningen och en mer strömlinjeformad väg till den övre delen.

Sammantaget bidrar denna avhandling till forskningsområdet livsförloppsförutsägelser (*life course prediction*) genom att undersöka förutsägbarheten hos två olika arbetsmarknadsutfall och genom att lyfta fram vikten av att undersöka förutsägbarheten hos undergrupper. Dessutom redogör den för olika sätt på vilka befolkningsregisterdata kan analyseras med hjälp av maskininlärningsmetoder för att förbättra vår förståelse av människans livsförlopp.

List of studies

1: Kuikka, S. (2024). The (Un)Predictability of Early (Un)Employment: A Machine Learning Approach. Socius: Sociological Research for a Dynamic World, 10.

2: Kuikka, S. (2026). Unpacking the Unpredictability of Early Unemployment: Exposing the Blind Spots of a Machine Learning Model.

3: Kuikka, S. (2026). Diverging Paths to Economic Destination: Uncovering Heterogeneity in Early Life Conditions Across and Within Income Classes.

Introduction

Machine learning, a new era of social inquiry

Artificial Intelligence (AI) is arguably the most important technological innovation of the 2020's (e.g., Gates 2023). The rapid development of AI has revolutionized the way humans operate in society: how they search for information, work, and interact with each other. Applications have been implemented in both the private and the public sectors at a remarkable speed, and the collection and storing of big data have developed alongside the technology that allows for high computation. When it comes to societal development, AI is a double-edged sword: it has been predicted to both promote societal development and reduce inequity, as well as to do the opposite (FRA, 2022; World Bank, 2026).

Behind modern AI systems is a set of statistical tools, namely, Machine Learning (ML). As early as 1950, Alan Turing wrote his famous conceptualization of "thinking machines", elaborating on a vision where machines can learn, or be taught, to understand large bodies of information and to perform logical deduction based on that (Turing, 1950). He wrapped the ideas grounded in computer science and engineering in a philosophical thought experiment that he called the Imitation Game, where machines learn and produce output to the extent that they can imitate a human. Fast-forward 70 years into the 2020s, the mechanics of Turing's proposal have developed into applications over-arching almost all of society, through AI.

ML is a set of data-driven methods from the field of computer science that flexibly find associations from complex empirical data, combined with a mathematically optimized learning algorithm through which the data is processed (see e.g., Breiman 2001). Through the AI revolution ML algorithms have become mainstream in society, although applications were in place long before, in the form of producing weather forecasts as precisely as possible, deciding whether a candidate for a bank loan is likely to pay it back, or suggesting the next song or a video the consumer is likely to enjoy, to mention some. Data are collected historically, as well as often in real time, and further used to train an ML algorithm to understand the complex patterns in the data, to either: A) group similar information (unsupervised learning), B) produce novel content (generative AI), or C) make the best possible guess for future

unseen data (supervised learning). These tools present the foundations for all modern AI systems, how they are built, and how they operate.

While ML provides vast opportunities for understanding complex patterns, the academic social sciences have only started to integrate these approaches into their workflow in recent years. There are applications of quantitative analysis of text corpuses to find sociological patterns and meanings (e.g., Bonikowski and Nelson 2022), satellite image recognition to, e.g., analyze contextual effects such as poverty (e.g., Daoud, Conlin, and Jerzak 2026), and individual-level life course analysis aiming to predict a given life outcome (e.g., Salganik et al. 2020).

This shift has begun forming a new paradigm revolving around predictive and algorithmic modelling, as well as the integration of prediction alongside explanation is being called for in the literature (see e.g., Hofman et al. 2021). While ML tools are now widely used in society and available to anyone, the social scientific research field in academia has so far been underutilizing them (Verhagen, 2022). This lack is particularly visible in individual-level life course analyses using survey or register data. Quantitative social scientific research has historically relied on a frequentist statistical approach, where hypotheses are formulated based on theories and previous empirical findings and then tested statistically. In the field of sociology this dates all the way back to Durkheim's classical study on suicide (Durkheim, 1897/1952) and has since developed through work by, e.g., Lazarsfeld (1958). These tools typically rely on (log-)linear assumptions and are powerful methods when the associations in the data are not particularly complex. However, this is not often the case when it comes to social processes.

We know that many social processes are complex by nature, and individual histories, life course processes, and life events interact with one another. However, quantitative sociological research still heavily relies on (log-)linear regression methods, that rely on (log-)linear assumptions of the relationships in the data and often simplify the complexity the human life course holds. These models struggle to account for higher-order interactions beyond two-three variables, with the researcher having to particularly specify each one. Moreover, they often model average effects and do not account for heterogeneity on either the side of predictors or the outcome unless explicitly modelled. They are also not well suited for simultaneously modelling highly correlated variables. These are some of the ways in which ML tools can complement conventional quantitative tools. They are specifically designed to understand patterns stemming from complexity, providing a strong argument for integrating them into the social research workflow.

While ML tools are available to everyone, most development and research is still done in the private sector by technology companies. The interests of the

private sector often differ from those of the public or academic sectors. Yet, the public sector has started implementing machine learning into its own processes, although the potential risks of reproducing or increasing inequalities are recognized (Regeringskansliet, 2025; SCB, 2023). It is crucial to increase the volume of academic research in the area of social machine learning, to better understand for whom the models work and for whom they do not, and to contribute to building machine learning-based public systems that do not amplify or create inequalities.

The overall aim of this dissertation is to explore how machine learning tools can be leveraged in combination with Nordic population register data to improve our understanding of the link between life course processes and labor market outcomes. The population registers contain plenty of detailed longitudinal information on full national populations that follow each person from birth or immigration onwards. They are a rich source of life course information that is used extensively in research. By combining the flexibility of machine learning tools with this data source, I contribute to developing new ways of extracting more information from the human life course than before, and to analyzing this rich information. As such, the thesis both advances the usage of ML tools for social science and provides guidance for how ML tools can be used in practice, for academics and the public sector alike.

Life course prediction with population register data

This thesis conceptualizes *life course prediction* as a distinct research field at the intersection of the social and computer sciences, following the recent shift in academic literature towards applying machine learning tools to analyze longitudinal human life course data: predicting life outcomes. The theoretical anchor is the Life Course Framework, which has a multi-disciplinary history dating back to the early 1900's, with a more formalized shape taking place from around the mid-century onwards (G. Elder et al., 2003). Machine learning tools are providing a new methodological approach for studying the human life course, with the goals being: A) further understanding human life course patterns and the sub-group heterogeneity within it, B) exploring whether the theories used for a given process are comprehensive, and C) understanding the limits of predictability of the human life course.

Life course prediction typically utilizes various ML tools and longitudinal data sources and focuses on one or several outcomes of interest to the social sciences. It stems from the ideas of further understanding the social patterns that so far have been studied mainly using conventional regression tools, the possibilities to improve designing public policies and targeted interventions, and the potential for theoretical and methodological development. The approach typically relies on thinking about the human life course as a Book of

Life—an individual script that each person builds throughout their lives—originating in the concept envisioned by Halbert Dunn (1946), and reimagined by Verhagen et al. (2025).

Some of the most famous examples in the line of life course prediction include the Fragile Families Challenge (Salganik et al., 2020), the Predicting Fertility Challenge (Sivak et al., 2024, 2026), and the life2vec project (Savcisens et al., 2023). What all these projects have in common is that they approach a social scientific research question with interdisciplinary framing, combining, e.g., sociology, demography, public health, computational social science, and computer sciences. The idea behind all these approaches is that the human life course can be best understood when looking through a multi-domain lens. They all incorporate the idea that when we have large-scale longitudinal data and statistical models that can capture complexity, it can improve our understanding about the social world.

The Fragile Families Challenge (Salganik et al., 2020) is generally considered as the first large-scale example in the field of life course prediction, and has often served as a benchmark to studies following it. In the context of the U.S., they leveraged birth cohort survey data from the Fragile Families and Child Wellbeing study from ages 0–9, in combination with machine learning tools, to predict six different life outcomes at age 15: household material hardship, child GPA, child grit, household eviction, primary caregiver participation in job training, and primary caregiver layoff. Furthermore, the study was set up using the common task method, where 160 interdisciplinary teams built their models independently from one another, to answer the same questions, using the same data. The benefits of the common task method are that it controls the risk of overfitting, as well as researcher degrees of freedom in, e.g., operationalizing the data. Their findings conclude that predictions for each outcome improve compared to the regression model benchmarks, but only marginally, some of them more than others. The discussion following the publication of this seminal work has lifted up points such as the importance of out-of-sample validation in the social sciences regardless of the methods used, as well as the potential the predictive paradigm may still hold if explored further, particularly related to various data sources, such as administrative population registers, and sub-group heterogeneity (e.g., Garip 2020).

In 2023, Savcisens et al. conducted a study where they used Danish population register data and a word embeddings algorithm to predict a range of outcomes such as early mortality and personality nuances. Their model outperformed other state-of-the-art models: for example, early mortality prediction improved greatly from the life table model used as a benchmark. Their study showed promise in using the combination of population registers and machine learning. In 2024, Sivak et al. designed and ran a mass collaboration challenge similar to the Fragile Families Challenge, but studying fertility outcomes,

more specifically, whether someone will have a child within 3 years from the observation period. They used both panel data as well as population registers from the Netherlands to conduct the study. The results highlight that the predictability of fertility in general is limited, and that the panel data fared better in the task than the registers (Sivak et al., 2026).

The aftermath of these seminal studies has also produced a new aim of uncovering unpredictability in the human life course (e.g., Lundberg et al. 2024). The core idea is to chart unpredictability and where it stems from. Lundberg et al. (2024) are studying the case of the Fragile Families Challenge and distinguish between reducible uncertainty—related to having access to better data or better algorithms—and irreducible uncertainty—randomness, human agency, and complex interactions. They also make an important point about what constitutes as a "good" or "bad" prediction: a social scientific prediction result could be benchmarked to the theoretical ceiling of what can be known, and not merely to a computer scientific ceiling of 100% accuracy. Hofman et al (2017), while discussing the relationship between prediction and explanation in human behavior, argued for transparency and standardization as additional markers for "good" prediction.

This dissertation has been written alongside the above-mentioned developments in the field of life course prediction. Given the calls for exploring administrative data sources for life course prediction, and the fact that ML models perform at their best with large amounts of detailed data, this dissertation uses Finnish and Swedish population register data and ML modelling together. Population registers are a longitudinal data source that follows each individual living in a country, from the moment they are born, or immigrated, until they die, or emigrate. These data record multi-domain information about the life course of full populations, that can be then linked together to follow each person's journey through life. Furthermore, these data are often linked across generations, facilitating the study of intergenerational processes. A multitude of research is done with population registers (using conventional methods) specifically in the social and health sciences in the context of countries that have a history of recording quality register data, such as the Nordic countries and the Netherlands.

In terms of sociological theory development, life course prediction can be particularly helpful. Analyzing a social process grounded in theory using several models, including conventional regression models as benchmarks, and ML models as the comparison, we may find gaps in predictability depending on the model used (e.g., Salganik et al. 2020). If the gap between the performance of the conventional model and the ML model is minor or non-existent, it can verify the theory used as the basis for analyzing the social process at hand. If there is a large gap, and the ML modelling provides additional performance, it is likely that the theory used is missing out on some

important interactions that can be found using a flexible ML modelling strategy. Further analyzing the results from the ML model can help locate the processes or sub-groups where the predictability is larger or smaller, thus providing an area for future social inquiry, and theory development. Another approach to theory development stems from the observations of large distances even between the best ML predictions and the perfect prediction. This observation has brought focus to disentangling from where this distance stems, with Lundberg et al. (2024) discussing in terms of missing data/theory vs. inherent randomness in the human life course, as well as defining a theoretical ceiling for how much of a given outcome or process we can predict, instead of assuming a ceiling of 100%. Furthermore, Yan and Rahal (2025) continue this by arguing that the inherent randomness in life course processes may just be something we have not learned how to measure yet. Finally, van de Rijt et al. (2026) argue that much of the distance can stem from unequally distributed luck in the population. All these discussions have stemmed from the emergence of life course prediction, and further research in the field can hopefully provide some answers.

This dissertation aims to explore not only overall predictability of labor market outcomes, but also sub-group predictability. Benchmarking all ML models to a logistic regression, the interest is founded on many dimensions of predictability. Through this, my work links to the practice of explainable AI (XAI), where the aim is to not merely predict an outcome, but to understand the predictability of a given outcome, and what builds it.

Aims

The aim of this dissertation is four-fold. First, I aim to contribute to the core paradigm of life course prediction, by answering how predictable the human life course is in terms of labor market outcomes, specifically long-term unemployment in early adulthood (study 1) and lifetime disposable income (study 3), in the Nordic context. Given that the field is emerging, this dissertation makes a novel contribution to this core question and provides guidelines for future research in the field.

Second, I aim to improve our understanding of whether some groups are more predictable than others and discuss what we can learn from differences in predictability across groups. I study this both in terms of origins (study 1), and outcomes (studies 1 and 3). This is important for two reasons. First, the answers contribute to the academic literature exploring how the life courses of different groups of people are connected to these labor market outcomes. Second, the answers provide essential information for those designing predictive tools that either facilitate or replace human decision makers in public agencies.

Third, I leverage a two-stage analytical setup, combining both supervised (SML) and unsupervised (UML) machine learning tools. Here, my first sub-aim is to understand the predictability of a given outcome using SML (studies 1 and 3), and second, to open up the prediction models to discuss which combinations of predictors are the ones that together contribute to the predictions using UML (studies 2 and 3). UML offers a toolset to map out the high dimensionality that the population registers present, as well as explore the results of the SML prediction models, in terms of the patterns that correctly or incorrectly predict the outcome (study 2), or that reveal the predictor heterogeneity across and within the outcome classes (study 3).

Fourth, I aspire to provide answers and guidance as to how data-driven ML methods can complement conventional methods in the study of the human life course, and present new ways of extracting details and yielding new information from the Nordic population registers.

Theoretical and empirical context

The two labor market outcomes of interest in this dissertation are long-term unemployment in early adulthood, i.e., having been unemployed consecutively for at least one year at age 25–30, and lifetime disposable income, i.e., a yearly mean income from labor and social subsidies, after tax, from age 18 to ~50. This section reviews relevant theories and empirical research on these two outcomes. This literature is later used to identify and discuss the mechanisms behind the labor market processes in question; to identify the most relevant and comprehensive predictor sets for the empirical studies; and to identify the gaps in the current theories worth exploring further.

Life course framework

The life course framework is a multidisciplinary perspective for understanding the human life course longitudinally through individuals' histories and future trajectories (G. Elder et al., 2003). It explores how and when life events unfold, how transitions from one stage of life to another happen, and how one's social relations are related to how their life course shapes out to be. The framework started developing during the early 1900's, but the mid and latter part of the century was when the framework established itself in the academic literature in the fields such as sociology, psychology, and history (Bynner, 2016). The institutionalization happened mostly through the development of research infrastructure required: developing and collecting longitudinal data sets, as well as technologies and methodologies facilitating the study of them. The life course framework can be viewed as a theory, leaning on Merton's (1949) conceptualization of the middle range theories, which calls for theories broad enough to work across several contexts but narrow enough to be testable with real data (e.g., Elder et al. 2003), although there are some that argue for a more traditional definition of a grand scientific theory, and therefore do not agree that the life course framework is one (e.g., Bernardi et al. 2019). Nevertheless, it is a framework particularly well-suited and far developed for understanding how multiple domains of human life jointly interact and unfold as a series of events, resulting in life experiences that sometimes vastly vary from one another.

One early conceptualization of the life course framework was written in 1946 by Halbert Dunn, who discussed the human life course in terms of the Book

of Life: how each person's life unfolds in particular ways from birth to death, and each event, interaction, and outcome across all life domains can be written into an individual book following that person's life. In his text Dunn was, in fact, writing about the importance of record-keeping and linkage, to be able to study Books of Life. In the aftermath of the Fragile Families Challenge, where machine learning tools were used to understand life course trajectories using survey data, the focus turned towards administrative population registers, and how they may enable new insights on life course studies with such methods. This is closely connected to the Book of Life idea, and one of the aims of this dissertation. The Book of Life approach has further been re-imagined specifically in this way by Verhagen et al. (2025), who aim to understand the lives of full Dutch populations using population registers from both social and health sciences, and analyzing them with advanced machine learning tools.

To focus on the outcomes in the studies included in this dissertation—long-term unemployment in early adulthood, and lifetime disposable income—I zoom into perspectives from within the life course framework. The outcome of long-term unemployment in early adulthood (studies 1 and 2) is inherently connected to the transition into adulthood, including, e.g., the transition from education to work life. While these transitions have historically happened rather standardly and densely with other changes (Rindfuss, 1991), there is a growing body of literature on the de-standardization of the transition into adulthood in the recent decades, particularly in the European context (e.g., Billari and Liefbroer 2010; Buchmann and Kriesi 2011). The Nordic setting is characterized by comparatively high labor market participation rates, standardized vocational educational systems, and family policies that to some extent mediate the gendered early labor market effects, but also paths of turbulent trajectories with patchy and unstable labor market attachment as well as social exclusion for some (Lorentzen et al., 2019). These characteristics can be further viewed through the lens of cumulative (dis)advantage (DiPrete & Eirich, 2006), which highlights how advantages and disadvantages accumulate over the life-course. This can result in a compounding of (dis)advantageous experiences affecting one's employability, and skills in landing and maintaining a job.

For lifetime disposable income, in addition to the transitioning into work life, and the (dis)advantages in chances for successful integration into the labor market, an important life course perspective is how a child's life chances are linked to the parents': the *linked lives* principle formulated by Elder et al. (2003). The main body of literature in income mobility discusses different mechanisms and extents that parental socioeconomic origins are connected to the child's income and economic opportunities (e.g., Blau and Duncan 1967; Solon 1992), while less of the literature leans on the life course framework, despite its seemingly natural connection. In addition to socioeconomic origins,

the larger context of life, including early life (adverse) experiences, are easy to link to how we should understand the accumulation of economic capital.

Different forms of capital

One key element to understanding the mechanisms behind both early life unemployment and lifetime disposable income lies in how different forms of capital accumulate and are transferred. I argue that at least five forms of capital play an important part for these outcomes: economic, human, social, cultural, and personal. Economic capital is intuitively connected with the labor market. On the individual level, it means that labor is exchanged for money, and money facilitates other things, such as making investments (Bourdieu, 1986). Human capital refers to productive capabilities, such as education and training, which again act as gatekeepers and facilitators for what type of work one lands (Becker, 1964). Social capital refers to connections, and is built through the networks one has access to, which again facilitate the kinds of work opportunities that open up (Portes, 1998). Cultural capital captures signaling and belonging and is typically built through ingrained habits. It includes, e.g., the type of education one has, language and dialect, and tastes, which all signal ones social position and suitability for a given position (Bourdieu, 1986). Personal or psychological capital refers to, e.g., mental resilience and attitudes and socialization toward work (Becker, 1996). It is linked to mental wellbeing and how one bounces back from adversity. Importantly, these forms of capital do not operate in isolation, but are in constant interaction with one another. By doing so, they compound into advantages and disadvantages in the labor market. As an example, economic capital facilitates purchasing time and services needed to acquire human capital, which again opens access to certain social networks and reinforces the cultural and personal capital needed to navigate the labor market. In the following sections I discuss how the forms of capital interact in the human life course, for each labor market outcome studied in this dissertation.

Long-term unemployment in early adulthood

In this section, I connect the life course framework and accumulation and transmission of different forms of capital to discuss the transition into the labor market, and how it can fail, leading to long-term unemployment in early adulthood. This is followed by a similar discussion about lifetime disposable income. This division is somewhat arbitrary, since many mechanisms are important for both outcomes, but it serves as a way to structure the thinking around the similarities and differences of the processes leading up to these two outcomes.

Transitioning into the labor market is an important life course marker in the youth that happens during a demographically dense period in life, where multiple transitions and paths in different domains overlap (e.g., Shanahan 2000). According to the contemporary view, these should not be treated as discrete events, but as a part of a biography shaping one's life course trajectory (e.g., Stauber, Walther, and Settersten, 2022). Transitioning from education to work is thus a puzzle with multiple simultaneous processes. To add to the palate, these trajectories start forming already at an early age, making the multigenerational family and upbringing environment crucial components in later work-life transitions (Caspi et al., 1998; Gilligan et al., 2018). When the early life provides adverse experiences, they can snowball into cumulative disadvantage (DiPrete & Eirich, 2006). Alternatively, positive experiences and lack of adversity can accumulate into more opportunities and positive outcomes (R. K. Merton, 1968). Thus, early life economic and social conditions accumulating act as a mechanism through which (dis)advantages accumulate in later life (O'Rand, 2006).

Acquiring human capital is important for employability as training is needed for most type of work (Becker, 1964). Early adversity in the form of limited economic capital in the family reduces the opportunity for formal or extracurricular academic training, that can help obtain higher or specialized education. It is also connected to housing quality and selecting the neighborhood to live in, and through that the available social networks. Furthermore, the interplay of disadvantaged human and social capital within the family can result in parents not having enough resources to invest in the child's development (Coleman, 1994). Poor economic capital also brings about family stress, which influences parenting and through that the child's life outcomes (G. H. Elder et al., 1994). Family stress can also influence the child's personal capital through reduced mental health and family coherence, which again influences academic success, anti-social behavior, and deviance (Conger, 1994; Sampson & Laub, 1994). Different forms of capital are in constant interplay with life course processes affecting risk of unemployment.

Lifetime disposable income

In this section I connect the life course framework and accumulation and transmission of different forms of capital to discuss the processes of accumulating income over lifetime, paying particular attention to how early life markers are connected to one's income level. Early life conditions in terms of, e.g., parental socioeconomic status, family constellation, migration status, urban/rural area of residence, are known to influence one's income trajectory (DiPrete & Eirich, 2006; Galster & Sharkey, 2017; McLanahan, 2004; Portes & Rumbaut, 2001; Roemer, 1998; Tickamyer & Duncan, 1990). The lives of family members, neighbors, and colleagues, and other social connections are

linked to one another in multiple ways, thus making central the life course framework principle of linked lives (G. Elder et al., 2003). The core idea is that the people a child lives around shape the child through their resources, constraints, and biographies: sons accumulate income in similar levels and patterns as their fathers (Cheng & Song, 2019), children with single or adoptive parents have lower educational attainment and engagement and poorer economic outcomes (Cavanagh & Fomby, 2012; Lindblad et al., 2003), migration background can lower educational outcomes and entry into the labor market through ethnic penalties and discrimination (Bursell & Bygren, 2025; Heath et al., 2008), the resources available or lacking in the neighborhood one lives in can support or hinder future economic outcomes (Brandén et al., 2023), and parenthood translating into economic penalty for mothers and premium for fathers (Angelov et al., 2016).

Parental socioeconomic status is the most well-studied condition for a child's socioeconomic destination, covering the scholarships of social stratification, social reproduction, and intergenerational mobility (e.g., Blau and Duncan 1967; Erikson and Goldthorpe 1992; Solon 1992). Economic capital accumulates through paid labor, where the processes of finding and maintaining work as described in the previous section about unemployment apply (Bourdieu, 1986). Intergenerational transmission of wealth is another channel for economic capital to accumulate for some, providing unequal opportunities. Wealth also acts as a safety net for making economic decisions including high risk and high reward. Parents with more economic capital can invest in the child's accumulation of human capital (Becker, 1964). Furthermore, parents' social capital is linked to the child's human capital: if parents with low income live and work in proximity to people similar to them, their social networks tend to form with similar people, and the opportunities for their children to socialize in ways that promote access to information about high paying jobs and education are limited (Coleman, 1988).

In addition to parental socioeconomic status, several other life course aspects can be viewed through the lens of accumulating and transferring capital. International migration plays into the early life conditions that form income trajectories in various ways (Portes & Rumbaut, 2001). Economic and human capital may travel with a migrant, and high levels of these can help individuals to end up on a positive income trajectory in the host country. Social and cultural capital are more difficult to transfer, and networks, as well as language skills and cultural dispositions often need to be rebuilt from scratch. Through the second demographic transition diverging family types have come into focus for how income trajectories are formed (McLanahan, 2004). Economic and social capital as well as stress in the family can accumulate in different ways for various family types. Rural and urban regions of residence provide different resources and opportunities for accumulating economic and human capital, with higher education being centered in larger cities, and internal

migration away from the family is required for those striving for it (Galster & Sharkey, 2017). Finally, gender is known to influence individuals' income trajectories. In the Nordic dual-earner model, occupational sorting, motherhood penalty, and fatherhood premium in the labor market yield differing economic capital for women and men (Mandel & Semyonov, 2006). Gendered occupational sorting is a process where human capital accumulation is not unequal vertically but is so horizontally. This is also reproduced through social capital, the networks that individuals have access to, the opportunities that open through them, and the type of education and work one is drawn to. All these aspects together wrap into forming one's income trajectory.

Methods and data

On statistical modelling cultures and sociological explanation

While most of the quantitative social sciences lean on frequentist statistics, i.e., the framework relying on p-values and confidence intervals to test hypotheses, this dissertation instead applies a toolset from computer science, namely, machine learning (ML). The logic of ML (for social sciences) is based on the idea that associations in large amounts of data can be extremely complex, and that pre-specifying a statistical shape, or functional form, can be misleading, and is often too simple for capturing the complexity of the data (Hastie et al., 2009). While the frequentist approach pre-defines a statistical functional form, such as logistic regression, and fits the data over it, the logic of ML is to let the data "speak" and reveal the associations within. In practice, this is often done by relying on a given algorithm, which pre-specifies some rules the data complies with, after which the complex associations in the data are explored.

The two modelling cultures in quantitative social science discussed here—frequentist statistics and ML—solve problems in fundamentally different ways. This relates to how social science understands what it means to explain the world. Starting from Durkheim's ([1897] 1952) idea of using statistics to establish social facts, by mid–1900's the statistical modelling of sociological processes developed into a variable-based modelling culture, where social attributes were operationalized into numerical objects and controlled for in a regression model, so as to establish relationships between X and Y that may operate through Z (Lazarsfeld, 1958). This view has since been the dominant approach in quantitative social sciences. Lately, this philosophical view has seen critiques. Abbott (1988) argued that human life and social processes do not behave like a linear equation, but that social reality is characterized by threshold effects, complex interactions, and temporal sequencing. He also critiqued the violation of the assumption of independent observations central to making frequentist causal claims: relying on most social theories, individuals cannot be assumed to be independent from one another. His critique was aimed at the core of the philosophy of social science. Later, Sørensen (1998) added to this critique by arguing for a missing mechanism link. His critique was that linear statistical models are disconnected from

sociological theories, and that their variables are not representative of people who act within social structures. If a sociologist theorizes about, e.g., class conflict, the theory does not suggest that the relationship between the variables is linear. He wanted to build formal mathematical models that specifically represent a given social mechanism: one of the foundations of the analytical sociology movement. Bayesian statistical culture, where probability distributions are a core concept, is a common tool in this branch of sociology.

Methodologically, the philosophical and mechanistic critique of variable-based sociology can also be bridged to computer science. Breiman (2001) made a distinction between the two modelling cultures—frequentist and ML—by stating that they disagree on how to solve the black box between X, the input variables, and Y, the output. While the goal of frequentist statistics is to understand how the variables relate to each other, the aim of ML is to accurately guess Y given an unseen set of X. Breiman critiqued conventional statistics for only relying on predefined functional forms to solve complex real-world problems—similarly to the social scientists' critique—as well as fitting one set of data on a function, instead of following the ML logic of training a model on one set of data and validating its performance on another, unseen set of data, a process known as cross-validation. This methodological critique has lately become visible in the social sciences. Hofman, Sharma and Watts (2017) argue that social scientists overly emphasize the importance of interpretable causal mechanisms and ignore whether the models can predict the outcomes of interest. Similarly, Berk (2006) critiques social scientists relying on simple functional forms, while social systems and human behavior are characterized by non-linearities and complex interactions. Mullainathan and Spiess (2017) argue that many vital policy questions require predicting an outcome and cross-validating it with unseen data. Finally, a group of these prominent critics published a plea to integrate explanation—the frequentist approach—and prediction—the ML approach—to complement each other and to further understand the complex social world (Hofman et al., 2021). This dissertation contributes to the underutilized part of this combination—prediction—to explore its possibilities and to guide further explanatory modelling.

Machine learning

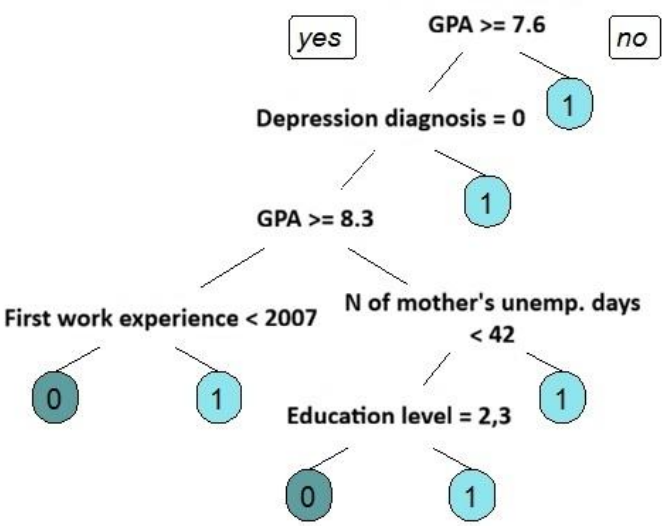
This dissertation can be seen as a methodological exploration of ML methods for life course sociology and labor market processes. I use several ML algorithms to predict outcomes, as well as to unpack the predictability of the populations as well as subgroups within. ML can be divided into two branches, supervised (SML), and unsupervised (UML), depending on their functioning logic and what type of information one wants to extract. While SML and UML are designed to perform separate tasks, in this dissertation I use them as a two-

stage methodological pipeline. I first use SML, more specifically decision tree-based models, to understand how well I can predict an outcome (long-term unemployment in early adulthood, binary classification; lifetime disposable income, 5–class classification), given the data from the population registers. This step maps high-dimensional inputs to find complex empirical patterns that predict the outcome and isolates the predictive signal from the empirical noise. The model outputs evaluation metrics that reveal how well this outcome can be predicted: A) on the macro level, and B) for each outcome class separately. I then move on to using UML to substantively elaborate on the SML findings post-hoc.

Supervised machine learning

The section above about statistical modelling cultures has mostly discussed SML. SML is intuitively close to the variable-based sociology approach of Lazarsfeld (1958), in the sense that it operates on a dataset with a known X and Y. The aim is to predict Y in an unseen dataset containing X. In this dissertation (studies 1 and 3) the SML models I use are classification trees (CART), Random Forest (RF), Gradient Booster (GB), and Extreme Gradient Booster (XGB). They all operate on the same foundational structure, the decision tree. The decision tree learns the input data through the logic of recursive partitioning, by continuously slicing the data to form tree branches, until finally reaching "leaves" that represent a group where most data points share the same outcome. See Figure 1 for an example of a decision tree. In this example the outcome is long-term unemployment, and the leaf nodes represent either yes (1) or no (0). The tree begins with the full sample at the top, at the root node. To make the first split, the algorithm searches through all the Xs in the data for the one that can best split the sample into two groups regarding the outcome and finds the best threshold of the variable to split at. The best split is often decided by the Gini Impurity measure, searching for the split that maximizes the number of people that share the same outcome on each side of the split. After the first split, there are now two samples, and the process of splitting continues. The tree forms branches, where each branch represents a combination of conditions or attributes that jointly form a given prediction. The terminal node, or leaf, is formed when the algorithm reaches a certain overfitting threshold, controlled by the researcher through setting hyperparameters for the algorithm. The branches are often asymmetrical, illustrating how social reality is highly interactive and non-linear.

Figure 1: A decision tree predicting long-term unemployment, from Kuikka, S. (2020).



The SML algorithms used in this dissertation are classification trees (CART), Random Forest (RF), Gradient Booster (GB), and Extreme Gradient Booster (XGB). Each of them leverages the decision tree as their base but vary in terms of how it is applied to reach the final predictions. CART builds a single tree, after which each person in the unseen holdout dataset is evaluated through how their attributes match the built tree, resulting in a predicted value of Y for the person. The predicted values of Y for all people in the holdout dataset are then aggregated to produce evaluation measures of how the algorithm fares in finding the complex patterns for the given social process or outcome. Given that one single tree can be very unstable, RF instead builds hundreds of decision trees independently and in parallel from random subsets of the data. To reach a final prediction for each person in the holdout data, the predictions from all the trees are averaged for a more stable result, reducing the variance and overfitting of a single CART. While RF uses a crowd wisdom logic, GB is designed to learn from mistakes already during the process of building the trees. GB builds several trees, but does so sequentially, with the goal of correcting errors in the previously built tree. The result is a sum of the result of each of the trees built, expressed as log-odds, and shrunk using a small multiplier. This is further transformed back into a probability space, giving a result comparable to the results from, e.g., CART and RF. Lastly, the XGB optimizes the computationally heavy GB algorithm to work in a parallel space using several computer cores, thus facilitating processing larger datasets faster, but fundamentally using the same logic of learning from mistakes as

GB. All the SML algorithms in this dissertation are used for classifying tasks and evaluated through a set of measures including accuracy (the proportion of correct predictions) and false positive/negative rate (the proportion of those observed to belong to one outcome category but predicted into another).

Unsupervised machine learning

In addition to SML, another branch called unsupervised ML (UML) is used in this dissertation (studies 2 and 3) to unpack the prediction models and to substantively elaborate on the SML findings post-hoc. UML operates on a set of known X's, but Y is not known. These algorithms instead identify the shape of the data, using a distance metric or variance maximization, and try to determine Y for each person in the data. The researcher then elaborates on what Y is, or what connects the people in each group, by substantively interpreting the combinations of X. I use two distinct UML algorithms in this dissertation: Principal Component Analysis (PCA) in study 2, and Hierarchical Agglomerative Clustering (HAC) in study 3.

PCA is used in study 2 to further understand the SML results for long-term unemployment in early adulthood in study 1, and why the models make many mistakes on one outcome class (unemployed) but perform well on the other (employed). PCA's goal is dimensionality reduction. This means that the algorithm aims to simplify the signal from tens of variables that are highly correlated and may measure the same concept into a smaller set of new variables, namely, principal components. It does so by mathematically squashing the variables together and defining principal components sequentially. The first component captures as much variance as possible, the second one as much of the remaining variance as possible, and so on. The resulting components are completely uncorrelated with each other, and form the output, or Y, of the model, which the researcher then interprets.

In study 3 analyzing lifetime disposable income, I use HAC to further understand what differentiates the people who end up in different parts of the income distribution, as well as the sub-groups within each income class. The goal of HAC is to group variables together into clusters where individuals are like one another. It does this by treating each person as their own cluster, calculating a mathematical distance between every single person, and merging the closest pair. The algorithm repeats this until everyone is clustered and outputs a diagram for the researcher to choose the number of clusters based on a set of evaluation measures. The clusters then represent the Y, specific sub-groups of people.

Summary of the models

Table 1 summarizes the classes of models I use in this dissertation, highlighting the methodological differences between frequentist statistics (here: logistic regression), SML (decision tree-based models), and UML (PCA and HAC). Logistic regression is used as a benchmark model for SML, to anchor the SML predictions into a baseline familiar to social sciences. Logistic regression represents the modelling culture where a functional form is preset, and which isolates an average independent effect of a single variable all else equal. In contrast, ML approaches the interactive complexity of the human life course. SML identifies the non-linear "recipes" of variables that maximize predictive accuracy. UML instead ignores the outcome and condenses the high-dimensional set of variables, revealing the latent structures present within the data.

Table 1: Summary comparison of the classes of models used in this dissertation

Feature of the model	Logistic Regression (Benchmark)	Supervised ML (Decision Tree-based)	Unsupervised ML (Clustering/Dimensionality Reduction)
Primary Logic	Hypothesis-driven: tests for a pre-defined relationship between X and Y	Prediction-driven: searches for empirical patterns that optimize predictive accuracy across the sample	Structure-driven: identifies latent structures, groupings, or lower-dimensional projections within the Xs
Primary Goal	Estimating average effects	Maximizing predictive power or minimizing loss	Discovering hidden structures
Questions	"How is X associated with Y?"	"How well can we predict Y? "Who is at risk of Y?"	"What types of lives/profiles/sub-groups exist?"
"All else equal" assumption	Fundamental: isolates the marginal effect of one X while holding others constant	Contingent: effect of Xi depends entirely on the values of other variables in the tree branch	Systemic: rejects the premise, predictors are a cohesive system
Functional form	Assumes a linear relationship between log-odds and predictors	Non-parametric, fits complex "step-functions" to the data, uses hyperparameters	Identifies the "shape" of the data using a distance metric or variance maximization
Linearity	Fixed: requires manual transformation for non-linear effects	Agnostic: does not assume linearity, but approximates both linear and non-linear effects	Varies: PCA assumes linear relationships, HAC relies on geometric distance instead.

		across ranges or categories of predictors	
Predictor relationship	Additive: assumes X1 and X2 are independent unless manually specified otherwise	Interactional: assumes the meaning of X1 is defined by its context with X2 and X3	Relational: X1, X2, and X3 covary to create a "type" or "dimension"
Inference goal	Isolating the signal: finding the unique contribution of X	Capturing the recipe: finding the specific combinations of X1, X2, and X3	Mapping the space: finding the underlying landscape
Interactions	Must be manually specified	Discovered automatically from the data	Captured implicitly, expressed as the "shape"
High dimensionality	Struggles: Prone to multicollinearity and over-fitting with too many predictors	Thrives: Feature selection is built in, can handle hundreds or thousands of predictors	Simplifies: designed to condense noisy high-dimensional data
Rare events	Sensitive: heavily biased toward the majority class	Robust: can isolate rare cases without skewing the model, through hyper-parameter tuning	Revealing: identifies rare types that averages would hide

Shapley values

To open up the black box of a SML model, the contemporary standard approach is to calculate Shapley values. Shapley values are a mathematical concept grounded in Game Theory, considered to be a core solution concept to understand the outcome of a game (Shapley, 1953). Nowadays Shapley values are being applied in both the fields of econometrics and AI.

The ML implementation of Shapley values is commonly SHAP (SHapley Additive exPlanations) (S. M. Lundberg & Lee, 2017). In the context of a SML model, the game is one single prediction, the players the variables X1, X2, and X3, and the goal is to determine how much credit or blame to give to each player for driving the prediction to where it landed. The calculation of a Shapley value is done by deducting the average expected value from the final predicted value, and the calculations are made for each possible combination of X for each person found in the prediction model. The logic is purely additive, making it a great complement to the black box of SML models.

For the purpose of life course research, SHAP provides the local interpretability of every single person in the dataset, connecting it back to the Book of Life approach (Dunn, 1946). However, sociologically it is important to note that Shapley values are calculated based on the obtained predictions regardless of whether the prediction is correct, close to correct, or far from correct. When implementing such tools in the social sciences where the accuracy of predictive models tends to be far from 100% (e.g., Kuikka 2024; Salganik et al. 2020), it is important to think about what Shapley values tell us and how we interpret them.

In this dissertation I used SHAP in study 3, in between the SML and UML approaches. I first predicted the outcome—lifetime disposable income—using several SML algorithms. I then chose the best performing SML model, calculated Shapley values for each X in the dataset, and plotted them globally to understand how much each social concept measured contributes to the predicted outcome. Furthermore, I used a UML approach, namely HAC, to group the Shapley values and to understand what types of sub-groups end up in each income class. In this stage, I specifically chose to plot only those correctly predicted in their respective income class, to be able to substantively discuss the social groups for whom the income class is predictable. This is an adjustment made from the perspective of a computational social scientist: while I choose not to interpret the incorrectly predicted cases, I allow for more room to substantively understand the social groups for whom lifetime disposable income is, in fact, predictable.

General remarks on life course prediction

In this dissertation I use ML algorithms which are not pre-trained, but which I train and validate myself on tabular longitudinal data from population registers. This means that none of this work relies on pre-trained models, such as Large Language Models (LLM). It is important to make this distinction given the criticism of pre-trained models stating that the results may vary heavily depending on the data used to pre-train the given algorithm, and that pre-trained models, which often are developed by private companies, can contain values that do not necessarily align with those of the research field (see e.g., Nelimarkka 2026). By training the algorithms myself, I have full control of the data being used for training. Furthermore, by using similar data for training and validation, I can better ensure that the results reflect the logic of the human life course learned from the training data, and that distribution shifts between the training and validation data do not bias the results.

Given that training and validating ML models, as well as performing further calculations such as Shapley values, is computationally expensive, it is important to motivate the use of additional computational resources compared

to the much lighter regression tools standard in the social sciences. While conventional parametric tools often aim to be parsimonious, the flexible ML models can provide more precision and better reflect the theories about the social world. The capacities of ML can be of particular importance for many policy questions, that aim to understand structures of inequality more comprehensively, and design effective interventions. Flexible ML models can find intersections of social attributes that would be difficult to find with conventional tools and thus allow for advancing our understanding about the inequalities or opportunities that accumulate on some. When a given outcome is costly either economically or socially, it is well worth spending computational resources for prevention as opposed to funding recovery from difficult social situations, such as unemployment.

Population register data

The Nordic countries have a long history of maintaining population registers that span over decades, containing individual-level information about the human life course, and covering several areas of life course information, including demographic, economic, and health data (Van Der Wel et al., 2019). This data source has been utilized in social research for decades to shed light on how Nordic people live their lives, how populations are shaped, and what types of challenges or social inequalities individuals and societies face. Most often this research is conducted with conventional statistical methods, and studies are often designed by following formats made for survey data use.

Population register data, compared to survey data, contain some particular features. They cover full populations, with new individuals born or emigrating into a country registered into the databases. They link each individual across multiple registers with a unique identifier, which for research purposes is transformed into a pseudo-ID (Ludvigsson et al., 2009). This enables following each person throughout their lives across different register holders. Each person is also linked to their family members, which enables studying intergenerational linkages. This facilitates, e.g., studying full birth cohorts in an intergenerational and longitudinal way (Paananen & Gissler, 2012).

The Nordic population registers offer a unique Book of Life view of the socio-demographic-economic life course of every person living in a country. This is a particularly detailed set of empirical data about people's lived lives, through the lens of administrative record-taking. In the population registers, lives unfold in highly specific bundled configurations. However, ways of extracting information from the population registers still mostly rely on tools that simplify the complexity typically through one estimand and often describing the average population. This washes away considerable heterogeneity, complexity, and detail about the way human lives unfold.

In recent decades, due to technological developments, the potential for analyzing big data has grown. We can now leverage more data in our analyses, use more complex methods, and describe big data in ways that were not previously possible. These technological developments for using more data and more complex methods of analysis also facilitate more detailed analysis of population registers. Combining these data with computational methods such as ML can facilitate further understanding of the human life course and can help extract more detailed information from the population registers than before.

These types of data are not, however, without limitations. The original purpose of population registers is book-keeping for the state, not research purposes nor understanding theoretical constructs (Van Der Wel et al., 2019). While they display a detailed longitudinal Book of Life view, this view excludes some essential elements for understanding humans as social actors: attitudes, emotions, motivations, and beliefs. Furthermore, given that data collection happens through varying state institutions, they are vulnerable to systemic reporting biases. These can stem from, e.g., municipal or regional variations in reporting habits, or from the fact that in cases like social work or law enforcement the registers report use of such services, but not non-use: if a crime is committed, but never reported, it never appears in the registers (Rodermond & Rostami, 2024). While registers are often said to cover full populations, there are people missing from the registers, such as undocumented immigrants (Van Der Wel et al., 2019). On the other hand, the registers are also known to contain some over-coverage, meaning that people who have emigrated can still be present in the registers (Monti et al., 2020). Simultaneously, while being impressive in how long back the data reaches, the definitions of each registered event may change over time, leading to inconsistencies in the definition of a given measurement. Finally, although the Nordic countries share similar features of welfare structures and register systems, conducting comparative research across countries is complex due to administrative bottlenecks and legal inconsistencies (Laugesen et al., 2021).

Finland

In studies 1 and 2 I leveraged population register data from Finland, more specifically the *1987 Finnish Birth Cohort Study* (Paananen & Gissler, 2012). This is a project that ran for two decades until the end of 2025, collecting data from each of the ~60 000 people born in 1987 in Finland and registered into the Medical Birth Record. The birth cohort was followed up until the age of 30, connecting information across 10 different population registers. Furthermore, the parents of the cohort members were followed up, facilitating not only longitudinal but also intergenerational research questions.

The Finnish Birth Cohort 1987 study is an example of a set of register data that are particularly broad and detailed. Information found in this study contains educational outcomes, labor market status, demographic life events, detailed health care data containing ICD codes, criminal records, social welfare interventions, and the timing of all these events in the life course. These types of data are particularly detailed, and ideal for doing multi-disciplinary life course research.

From these data I operationalized a dataset with the goal of understanding the complex life course structures that together predict long-term unemployment in early adulthood. The outcome was operationalized by being registered as a job seeker for more than 365 consecutive days during ages 25 and 30. It aims to capture individuals at an age where most have finished their education and have transitioned or are transitioning into the labor market. The predictor variables were chosen and operationalized based on theories and previous research on the outcome. Some of them were operationalized cross-sectionally, such as grade point average at age 15, or highest education level attained before age 25. Some were operationalized continuously, such as mother's highest education level, some as binary, such as ever had a criminal conviction. Some were operationalized yearly, such as the year of out-of-home placement, and some as to follow critical age groups regarding school transitions (0–6 years; 7–12 years; 13–15 years; 16–24 years), such as receiving mental health diagnoses, or experiencing parental death. Using flexible ML tools together with such wide and detailed data operationalized to follow the early life course facilitates finding complex interactions in the data that can predict the outcome in question.

Sweden

In study 3 I leveraged population register data from Sweden, obtained through the project *Intra-, inter-, multigenerational and categorical inequality* at Stockholm University Sociology Department. The foundational data source in this project is the Multi-generational register (Olausson & Lindberg, 2017). It contains people born after 1932 who have been registered to live in Sweden sometime since 1961. All people are also connected to their biological, and if present, adoptive parents. These data allow for a particularly long-term follow-up period, and in the case of study 3, measuring the income of the birth cohorts 1967–1971 throughout their working lives, as well as having access to socioeconomic markers for their parents.

From these data I operationalized a dataset to understand which early life factors, in addition to parental SES, predict one's lifetime disposable income. The outcome is lifetime disposable income, which includes both labor income

and welfare subsidies, presented after taxes. It is measured yearly from age 18 onwards, up until the year 2018 (47–51 years old), and indexed to the year 2000. Therefore, it aims to measure the amount of income that each person has at their disposal, throughout their life. I operationalized it as the yearly mean across three decades, excluding any years where a person did not have a measurement (due to, e.g., migration or death). I further operationalized lifetime disposable income into 5 categories based on observing the similarity of the income decile distributions. The predictor variables were based on theoretical and empirical literature on precursors for income, representing the early life conditions in terms of family socioeconomic situation, family composition, migration background, and urban/rural region.

The two data sources differ in some ways but complement one another. The Finnish data has a particularly wide and detailed set of life course information, including, e.g., International Classification of Diseases (ICD) codes for diagnoses, criminal records, and out-of-home placement. The Swedish data on the other hand has a particularly long follow-up time to calculate the outcome over a large span of working life. Together they paint a fairly comprehensive picture of the working-life course.

On research infrastructure

Combining population register data and computationally heavy ML methods presents some challenges. During the writing of this dissertation, I have had to make many compromises on my analytical strategy and research questions due to the limited infrastructure for secure processing of sensitive individual-level data. Here I will elaborate on the lessons I have learned, as well as the need to further develop the infrastructure to facilitate such analyses.

For studies 1 and 2 using data from Finland, the data were stored on internal servers at the Finnish Institute for Health and Welfare, with no additional computational processing capacity beyond that of a regular laptop. Given that the project was established more than 20 years ago and the processing infrastructure was set up to meet the needs of the day, by the time of this dissertation it was already outdated. This meant that I was very limited in choosing which ML algorithms to use when analyzing these data and calculating SHAP values was not an option given the constraints on computational power. This project ended in December 2025, and future projects are likely to be implemented within the FIONA remote access system operated by Statistics Finland (2026).

Access to the data for study 3 was through MONA, the remote access system operated by Statistics Sweden, like FIONA in Finland (Statistics Sweden, 2026). Both are platforms where individual-level population register data can

be accessed securely by a researcher through a remote desktop, while being stored and processed at servers internal to each respective institute. While these remote access systems have many advantages, such as centralized storage and access to detailed individual-level data, their resources were also designed for a time when computational requirements were not heavy: to this date neither MONA nor FIONA contain Graphics Processing Units (GPU) required for computationally heavy processing of large amounts of data, such as advanced ML for longitudinal data of full populations. With regards to study 3, MONA enabled me to use some more computationally heavy tools compared to studies 1 and 2, such as calculating SHAP values, but I continued to have limitations regarding the sample size to calculate them for. Particularly when using Random Forest, where deep trees are required for better accuracy for life course research, it simultaneously increases the number of SHAP values to be calculated exponentially due to the number of interactions.

To fulfill the need for more computational power, the National Academic Infrastructure for Supercomputing in Sweden has recently been connected to MONA, and researchers can apply for processing time from servers dedicated to processing sensitive individual-level data (NAISS, 2025). These servers are equipped with both CPU for lighter processing as well as GPU for heavier processing. This infrastructure is also developing constantly, as these sensitive processing servers are getting even more resources in late 2026 through transferring into the new Arrhenius cluster (Arrhenius, 2026). This supercomputer access was not yet applicable for this dissertation but is available for any future research.

To apply computational analysis methods in combination with population register data, it is imperative that the infrastructure for these analyses is developed. An example to follow in terms of how to develop the infrastructure for data processing, as well as collection and dissemination, is the Open Data Infrastructure for Social Science and Economic Innovations in the Netherlands, ODISSEI (2026). They have developed a national centralized system for accessing and securely processing the Statistics Netherlands' population register data through the ODISSEI secure supercomputer, as well as supporting services such as training in computational methods for researchers. In addition to the infrastructure needed for population register data, ODISSEI also coordinate the LISS panel data of Dutch households, which can be linked to the register data of the respective household panel members, resulting in a rich source of data to understand the human life course. Furthermore, ODISSEI provides a code library including sources for both the register and the panel data, as well as support in FAIR principles, ethical reflection, and grant application.

Van Der Wel et al. (2019) identified multiple bottlenecks in Nordic register data access, specifically when working on research questions with country

comparisons, due to the lack of unified systems of access to the data, as well as legislative differences across countries. The wealth of data the registers contain will not be of service to society until research processes are made smoother. The development of a secure infrastructure that facilitates cross-country comparisons can further advance science and our understanding of contemporary society.

Ethical considerations

The Swedish Ethical Review Act (Lag Om Etikprövning Av Forskning Som Avser Människor, 2003) states that the gain of knowledge through research done needs to outweigh any potential risks that come to the subjects of the study. When using population register data—detailed longitudinal information from full populations—this includes both ensuring the data is processed securely and anonymously and also reporting findings in a way that does not pose risks for any groups of people. Furthermore, in the case of processing sensitive personal information, as defined by the EU General Data Protection Regulation (GDPR), applying for an ethical approval from the Swedish Ethical Review Authority (Etikprövningsmyndigheten) is a requirement. Such data include, for example, information concerning one's health, and racial or ethnic origin.

This dissertation leverages population register data from two different sources and countries, with two different procedures of secure accessing of the data. Studies 1 and 2 are conducted using data administered by The Finnish Institute for Health and Welfare (THL). The data access was granted to the researcher through an internal review process. The data were stored at local servers at THL in Finland and accessed through a secure connection using a two-phase identification, a VPN, and a specific device (a laptop provided by THL). The data were delivered via the National Birth Cohort 1987 project, which ended 31.12.2025, and after which the data were archived and no longer accessible to researchers. These two studies included sensitive personal data and have received ethical approval (Dnr 2022–03844–01).

Study 3 leverages population register data from Sweden, through the project Intra-, inter-, multigenerational and categorical inequality, administered by the Department of Sociology at Stockholm University. The project has received ethical approval (Dnr 2023–00820–02). These data are stored by Statistics Sweden (SCB), and accessed through MONA platform, a remote desktop system provided by SCB, which the researcher accesses through a personal identification system.

In both data sources, each person in the data is given a pseudo-ID (PID) before researchers have access to the data, thus making sure that a researcher never sees the true personal identification numbers in the population. The PID is

crucial to have access to since it is used to link information associated with a specific person across different registers, to build longitudinal and multi-domain datasets, while preserving the anonymity of the subjects. While identifying individuals in the data is difficult, it can potentially be possible. This would, however, require active efforts from the researcher, and is neither allowed nor in the interest of sociological inquiry. To take further measures to safeguard against any misuse, I have engaged in ethical training, such as attending The Swedish Interdisciplinary Graduate School in Register-Based Research, which focuses on several aspects of register data use, including legal and ethical ones. Furthermore, before accessing either data source, I have signed legally binding documents to ensure I follow secure data management procedures and ethical conduct. Given all these measures, the risks that the research conducted in this dissertation poses for the subjects of the studies are minimized, and the potential gains for improving our understanding about the complexities on multi-domain life courses for labor market outcomes outweigh the potential risks. Moreover, since the analytical tools used in the dissertation—Machine Learning algorithms—are already used by public agencies to improve their decision-making processes, the improved understanding on how they operate, and for whom do they work better or worse, is essential. In presenting my results, I pay attention both to not enabling identification of particularly vulnerable groups, and to not label such groups in negative ways.

Summary of studies

Study 1 - The (Un)Predictability of Early (Un)Employment: A Machine Learning Approach

Transitioning into working life is an important marker in the life course, and people's trajectories start forming already from an early age. This study explores the extent to which one can predict who will end up unemployed at ages 25–30 based on what we know from their early life course, and whether there is variance in predictability across the outcome groups and socioeconomic background. The study focuses on the context of Finland, and leverages data from the full birth cohort of 1987 (analytical sample = 58,179). Theories and previous literature on unemployment are the foundation for the expected mechanisms, as well as choosing and operationalizing the set of predictors used in this study. The information on previous life course available covers a wide range of domains, from educational attainment, partnership and family formation, records of mental health, criminal records, out-of-home placement, parental socioeconomic status, and parental unemployment and death. Some of these are operationalized longitudinally, some cross-sectionally, and some following the ages of school transitions (0–6; 7–12; 13–15; 16–24). The study finds that the predictability of long-term unemployment improves with using decision tree-based machine learning algorithms (accuracy for CART 72%; Random Forest 73%; Gradient Booster 74%), compared to logistic regression (accuracy 65%), but that the algorithms miss-classify most of the unemployed net of baseline likelihood to be unemployed: using Gradient Booster, 74% of the unemployed are miss-classified as employed. These are important findings given that the unemployed is typically the group for which policies and interventions are targeted. For the low SES group, defined as low parental education, the overall predictability is worse than for the high SES group, although correctly predicting the outcome unemployment improves slightly. The results suggest that the current literature—which is what the choice of predictors was based on—may not yet fully understand the mechanisms underlying early unemployment.

Study 2 - Unpacking the Unpredictability of Early Unemployment: Exposing the Blind Spots of a Machine Learning Model

Building on the findings of study 1, study 2 aims to unpack these results to explore the combinations of early life characteristics that predict someone as employed or unemployed, and which combinations do this correctly and which incorrectly. Substantively this will give us an indication of which social mechanisms are the blind spots for the models. Using the same data as study 1 and the results from the best performing model (Gradient Booster), this study uses Principal Component Analysis to identify components of early life course characteristics that predict individuals as either employed (correctly vs. incorrectly) or unemployed (correctly vs. incorrectly). The components for the correctly predicted cases align with the theories and mechanisms described in the literature. For the group observed employed but incorrectly predicted as unemployed, the machine learning model is blind to the predictive effect of parental unemployment. For the group observed unemployed but predicted employed, family cohesion, mental wellbeing, lack of adversity, and father's high education level are the blind spots, potentially suggesting a joint resilience effect.

Study 3 - Diverging Paths to Economic Destination: Uncovering Heterogeneity in Early Life Conditions Across and Within Income Classes

Study 3 integrates the life course framework and machine learning tools into the study of income. The focus is on income trajectories formed by early life conditions, and the predictability of income class. The study also explores heterogeneity across and within income classes. The context is Sweden, and the study focuses on analyzing the birth cohorts of 1967–1971. Theoretical literature elaborates on the most well-studied predictor, parental socioeconomic status, in combination with other important domains of life, such as geographic origins, family structure, and gender. Using tree-based machine learning algorithms, the first stage explores the overall predictability of lifetime disposable income, measured by the yearly mean income over 3 decades. Compared to logistic regression, overall predictability improves when using machine learning algorithms. Out of all the income classes, the middle-income class has the highest predictive accuracy, whereas the lowest income class has the lowest. Using Shapley Additive Explanations, each predictor receives a calculated value for how much it contributed to the predictions. These values are further analyzed using Hierarchical Agglomerative Clustering to obtain clusters of early life conditions within

each income class, revealing groups of people who end up in the same income class but through different paths. These results highlight that the pathways to ending up at the bottom of the income distribution are complex and varying, whereas at the top of the distribution, the pathways are particularly streamlined.

Discussion and conclusion

The overall aim of this dissertation has been to explore how machine learning tools can be leveraged in combination with Nordic population register data to improve our understanding of the link between life course processes and labor market outcomes. The contributions are multi-fold. First, I have contributed to the emerging sub-field of life course prediction by studying the predictability of long-term unemployment in early adulthood in Finland and lifetime disposable income in Sweden. The results revealed that in both cases, predictive modelling identifies life course patterns better than conventional regression modelling overall, following the findings of, e.g., Salganik et al. (2020) and Savcisens et al. (2023). This provides strong incentive for integrating modelling practices that rely on prediction, as suggested by Hofman et al. (2021), in particular for life course research and the study of labor market outcomes.

Second, my contribution extends to highlighting that the predictability of the two labor market outcomes differs by social groups. My findings reveal that specifically the most disadvantaged groups—unemployed and low income class—are more difficult to predict than others. This is a novel contribution within the area of life course prediction, as most previous studies report only overall predictability for each outcome, not sub-group specific results (e.g., Salganik et al. 2020; Savcisens et al. 2023; Sivak et al. 2026). This contribution also relates to the work by, e.g., Lundberg et al. (2024) about understanding the origins of unpredictability. For academic research, these findings highlight that even in the case of such well-researched outcomes, using theoretically and empirically derived predictors, we do not necessarily understand the outcomes well at the more disadvantaged parts of the distribution. This calls for formulating both new empirical strategies as well as theoretical approaches. Furthermore, these findings have particular implications for public agencies, where machine learning based tools for aiding decision-making are already being applied (SCB, 2023), while research in understanding the predictability of these groups is lagging. To be able to design a predictive tool that aids in decision-making and that often targets the more disadvantaged groups, knowing how poorly these models can fare for those groups poses new questions regarding how to best develop and apply these tools.

Third, I have applied a novel, two-stage analytical strategy that leverages both supervised and unsupervised machine learning tools, first predicting the outcome at hand, and then unpacking the results and mapping out the high-dimensionality of the predictors. Through UML I have demonstrated: (A) which combinations of life course predictors correctly and incorrectly predict (un)employment (study 2), and (B) which combinations of early life conditions produce heterogeneity across or within income classes (study 3). This approach has highlighted the higher-order interactions of predictors behind these processes, thereby revealing that the theories we rely on are not overarching. Theories regarding unemployment may, in fact, mostly discuss employment, and theories about income accumulation vastly vary across low, middle, and high income. Connecting the theories that work for different parts of the distributions allows us to further develop the theories used for understanding the heterogeneity of these labor market processes.

Fourth, I have explored tools that complement the conventional methods, on how to extract information from the wealth of data that are the population registers. I found that ML tools can be powerful in further understanding the detailed ways in which the human life course unfolds in the population registers. However, I have discovered that to leverage these methods for a more comprehensive understanding of what the population registers can teach us about life courses, it is essential that we develop the infrastructure around it (see e.g., Van Der Wel et al. 2019). While conducting the analysis for this dissertation, I have constantly had to make decisions that limit the scope of discovery. The large amount of data, in combination with the methods that perform exponentially more calculations than the conventional ones, demands an environment where we can enable both scientific discovery and rigorous security for processing such detailed, sensitive, individual-level data (e.g., ODISSEI 2026). Developing these infrastructures is imperative for further discoveries in this line of work.

References

Abbott, A. (1988). Transcending General Linear Reality. *Sociological Theory*, *6*(2), 169. https://doi.org/10.2307/202114

Angelov, N., Johansson, P., & Lindahl, E. (2016). Parenthood and the Gender Gap in Pay. *Journal of Labor Economics*, *34*(3), 545–579. https://doi.org/10.1086/684851

Arrhenius. (2026). https://hpc.pages.naiss.se/training/selecting-modules/arrhenius/

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* (G. S. Becker, Ed.; 1st ed.). National Bureau of Economic Research.

Becker, G. S. (1996). *Accounting for Tastes* (1st ed.). Harvard University Press.

Berk, R. (2006). An Introduction to Ensemble Methods for Data Analysis. *Sociological Methods & Research*, *34*(3), 263–295. https://doi.org/10.1177/0049124105283119

Bernardi, L., Huinink, J., & Settersten, R. A. (2019). The life course cube: A tool for studying lives. *Advances in Life Course Research*, *41*, 100258. https://doi.org/10.1016/j.alcr.2018.11.004

Billari, F. C., & Liefbroer, A. C. (2010). Towards a New Pattern of Transition to Adulthood? *Advances in Life Course Research*, *15*, 59–75. https://doi.org/10.1016/j.alcr.2010.10.003

Blau, P. M., & Duncan, O. D. (1967). *The American Occupational Structure*. Wiley Books.

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. *Sociological Methods & Research*, *51*(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Bourdieu, P. (1986). The Forms of Capital. In *Handbook of Theory and Research for the Sociology of Education*. Greenwood Press.

Brandén, M., Haandrikman, K., & Birkelund, G. E. (2023). Escaping one's disadvantage? Neighbourhoods, socioeconomic origin and children's adult life outcomes. *European Sociological Review*, *39*(4), 601–614. https://doi.org/10.1093/esr/jcac063

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–215.

Buchmann, M. C., & Kriesi, I. (2011). Transition to Adulthood in Europe. *Annual Review of Sociology*, *37*(1), 481–503. https://doi.org/10.1146/annurev-soc-081309-150212

Bursell, M., & Bygren, M. (2025). The making of ethnic segregation on the labor market—Evidence from a field experiment. *European Societies*.

Bynner, J. (2016). Institutionalization of Life Course Studies. In M. J. Shanahan, J. T. Mortimer, & M. Johnson (Eds.), *Handbook of the Life Course Volume II*. Springer International Publishing AG Switzerland.

Caspi, A., Wright, B. R. E., Moffitt, T. E., & Silva, P. A. (1998). Early Failure in the Labor Market: Childhood and Adolescent Predictors of Unemployment in the Transition to Adulthood. *American Sociological Review*, *63*(3), 424–451.

Cavanagh, S. E., & Fomby, P. (2012). Family Instability, School Context, and the Academic Careers of Adolescents. *Sociology of Education*, *85*(1), 81–97. https://doi.org/10.1177/0038040711427312

Cheng, S., & Song, X. (2019). Linked Lives, Linked Trajectories: Intergenerational Association of Intragenerational Income Mobility. *American Sociological Review*, *84*(6), 1037–1068. https://doi.org/10.1177/0003122419884497

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *The American Journal of Sociology*, *94*.

Coleman, J. S. (1994). Social Capital, Human Capital, and Investment in Youth. In A. C. Petersen & J. T. Mortimer (Eds.), *Youth Unemployment and Society*. Cambridge University Press. https://doi.org/10.1017/CBO9780511664021

Conger, R. D. (Ed.). (1994). *Families in Troubled Times—Adapting to Change in Rural America* (1st ed.). Routledge.

Daoud, A., Conlin, C., & Jerzak, C. T. (2026). Chinese vs. World Bank development projects: Insights from earth observation and computer vision on wealth gains in Africa, 2002–2013. *World Development*, *202*, 107328. https://doi.org/10.1016/j.worlddev.2026.107328

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments. *Annual Review of Sociology*, *32*(1), 271–297. https://doi.org/10.1146/annurev.soc.32.061604.123127

Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health*, *36*, 1412–1416.

Durkheim, É. (1952). *Suicide, a study in sociology* (J. A. Spaulding & G. Simpson, Trans.). Routledge. (Original work published 1897)

Elder, G. H., Robertson, E. B., & Ardelt, M. (1994). Families Under Economic Pressure. In R. D. Conger (Ed.), *Families in Troubled Times—Adapting to Change in Rural America* (1st ed., p. 25). Routledge.

Elder, G., Johnson, M. K., & Crosnoe, R. (2003). The Emergence and Development of Life Course Theory. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the Life Course*. Kluwer Academic Publishers.

Erikson, R., & Goldthorpe, J. H. (1992). *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Clarendon Press.

FRA. (2022). *Bias in algorithms – Artificial intelligence and discrimination*. European Union Agency for Fundamental Rights. https://doi.org/doi:10.2811/25847

Galster, G., & Sharkey, P. (2017). Spatial Foundations of Inequality: A Conceptual Model and Empirical Overview. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, *3*(2), 1. https://doi.org/10.7758/rsf.2017.3.2.01

Garip, F. (2020). What Failure to Predict Life Outcomes Can Teach Us. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8234–8235. https://doi.org/10.1073/pnas.2003390117

Gates, B. (2023, March 21). The Age of AI has begun. *Gates Notes*. https://www.gatesnotes.com/the-age-of-ai-has-begun

Gilligan, M., Karraker, A., & Jasper, A. (2018). Linked Lives and Cumulative Inequality: A Multigenerational Family Life Course Framework. *Journal of Family Theory & Review*, *10*(1), 111–125. https://doi.org/10.1111/jftr.12244

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning—Data Mining, Inference and Prediction* (2nd ed.). Springer.

Heath, A. F., Rothon, C., & Kilpi, E. (2008). The Second Generation in Western Europe: Education, Unemployment, and Occupational Attainment. *Annual Review of Sociology*, *34*(1), 211–235. https://doi.org/10.1146/annurev.soc.34.040507.134728

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Kuikka, S. (2020). *Classification Tree Based Algorithms in Studying Predictors for Long-Term Unemployment in Early Adulthood—An Exploratory Analysis Combining Supervised Machine Learning and Administrative Register Data*.

Kuikka, S. (2024). The (Un)Predictability of Early (Un)Employment: A Machine Learning Approach. *Socius: Sociological Research for a*

Dynamic World, *10*. https://doi.org/doi.org/10.1177/23780231241286655
Lag Om Etikprövning Av Forskning Som Avser Människor, 2003:460, Utbildningsdepartementet (2003).
Laugesen, K., Ludvigsson, J. F., Schmidt, M., Gissler, M., Valdimarsdottir, U. A., Lunde, A., & Sørensen, H. T. (2021). Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries. *Clinical Epidemiology*, *Volume 13*, 533–554. https://doi.org/10.2147/CLEP.S314959
Lazarsfeld, P. (1958). Evidence and Inference in Social Research. *Daedalus*, *87*(4), 99–130.
Lindblad, F., Hjern, A., & Vinnerljung, B. (2003). Intercountry adopted children as young adults—A Swedish cohort study. *American Journal of Orthopsychiatry*, *73*(2), 190–202. https://doi.org/10.1037/0002-9432.73.2.190
Lorentzen, T., Bäckman, O., Ilmakunnas, I., & Kauppinen, T. (2019). Pathways to Adulthood: Sequences in the School-to-Work Transition in Finland, Norway and Sweden. *Social Indicators Research*, *141*(3), 1285–1305. https://doi.org/10.1007/s11205-018-1877-4
Ludvigsson, J. F., Otterblad-Olausson, P., Pettersson, B. U., & Ekbom, A. (2009). The Swedish personal identity number: Possibilities and pitfalls in healthcare and medical research. *European Journal of Epidemiology*, *24*(11), 659–667. https://doi.org/10.1007/s10654-009-9350-y
Lundberg, I., Brown-Weinstock, R., Clampet-Lundquist, S., Pachman, S., Nelson, T. J., Yang, V., Edin, K., & Salganik, M. J. (2024). The origins of unpredictability in life outcome prediction tasks. *Proceedings of the National Academy of Sciences*, *121*(24), e2322973121. https://doi.org/10.1073/pnas.2322973121
Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, *30*.
Mandel, H., & Semyonov, M. (2006). A Welfare State Paradox: State Interventions and Women's Employment Opportunities in 22 Countries. *American Journal of Sociology*, *111*(6), 1910–1949. https://doi.org/10.1086/499912
McLanahan, S. (2004). Diverging destinies: How children are faring under the second demographic transition. *Demography*, *41*(4), 607–627. https://doi.org/10.1353/dem.2004.0033
Merton, R. (1949). On Sociological Theories of the Middle Range. In *Social Theory and Social Structure*. Simon & Schuster.
Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Monti, A., Drefahl, S., Mussino, E., & Härkönen, J. (2020). Over-coverage in population registers leads to bias in demographic estimates. *Population Studies*, *74*(3), 451–469. https://doi.org/10.1080/00324728.2019.1683219

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

NAISS. (2025). *BIANCA*. NAISS. https://www.naiss.se/resource/bianca/

Nelimarkka, M. (2026). MarxistLLM: Fine-tuning a language model with a Marxist worldview. *Big Data & Society*, *13*(2), 20539517261447831. https://doi.org/10.1177/20539517261447831

ODISSEI. (2026). *About ODISSEI*. https://odissei-data.nl/about-odissei/

Olausson, P. O., & Lindberg, J. (2017). *Multi-generation register 2016 – A description of contents and quality* (Population and Welfare - Background Facts 2017:2). Statistics Sweden.

O'Rand, A. M. (2006). Stratification and the Life Course: Life Course Capital, Life Course Risks, and Social Inequality. In *Handbook of Aging and the Social Sciences* (6th ed., pp. 145–162). Academic Press.

Paananen, R., & Gissler, M. (2012). Cohort profile: The 1987 Finnish Birth Cohort. *International Journal of Epidemiology*, *41*(4), 941–945. https://doi.org/10.1093/ije/dyr035

Portes, A. (1998). Social Capital: Its Origins and Applications in Modern Sociology. *Annual Review of Sociology*, *24*. https://doi.org/10.1146/annurev.soc.24.1.1

Portes, A., & Rumbaut, R. (2001). *Legacies: The story of the Immigrant Second Generation* (1st ed.). University of California Press.

Regeringskansliet. (2025). *Ökat jämställdhetsperspektiv när myndigheter använder AI*. Arbetsmarknadsdepartementet.

Rindfuss, R. R. (1991). The Young Adult Years: Diversity, Structural Change, and Fertility. *Demography*, *28*(4), 493–512. https://doi.org/10.2307/2061419

Rodermond, E., & Rostami, A. (2024). Strengths and Weaknesses of Using Register Data to Study Extremism and Terrorism. *Studies in Conflict & Terrorism*, 1–16. https://doi.org/10.1080/1057610X.2024.2361949

Roemer, J. E. (1998). *Equality of Opportunity*. Harvard University Press.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Sampson, R. J., & Laub, J. H. (1994). Urban Poverty and the Family Context of Delinquency: A New Look at Structure and Process in a Classic Study. *Child Development*, *65*, 523–540.

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., Zettler, I., & Lehmann, S. (2023). Using sequences of life-events to predict human lives. *Nature Computational Science*. https://doi.org/10.1038/s43588-023-00573-5

SCB. (2023). *AI-användning i företag och offentlig sektor* (p. 13). Statistics Sweden. https://www.scb.se/contentassets/ea0e9cccd58343e7a07fe4c055f8fad2/nv0116_2022a01_br_nvftbr2301.pdf

Shanahan, M. J. (2000). Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective. *Annual Review of Sociology*, *26*, 667–691.

Shapley, L. S. (1953). A Value for n-Person Games. In H. W. Kuhn & A. Tucker (Eds.), *Contributions to the Theory of Games* (Vol. 2, pp. 307–317). Princeton University Press. (Original work published Princeton University Press)

Sivak, E., Cantrell, E. M., Emery, T., Garcia-Bernardo, J., Hafner, F., Karpinska, K., Lüken, M., Mendrik, A., Mulder, J., Ren, H., Satish, V., Verhagen, M., Mainieri, A., Pankowska, P., Abdel Ghany, J., Arpino, B., Cassani, G., Hellstrand, J., Ivanova, K., … Stulp, G. (2026). *Births are difficult to predict even with rich survey and full-population register data* (arXiv:2609.01194v1). arXiv. https://doi.org/10.48550/arXiv.2609.01194

Sivak, E., Pankowska, P., Mendrik, A., Emery, T., Garcia-Bernardo, J., Höcük, S., Karpinska, K., Maineri, A., Mulder, J., Nissim, M., & Stulp, G. (2024). Combining the strengths of Dutch survey and register data in a data challenge to predict fertility (PreFer). *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-024-00275-6

Solon, G. (1992). Intergenerational Income Mobility in the United States. *The American Economic Review*, *82*(3), 393–408.

Sørensen, A. B. (1998). Theoretical mechanisms and the empirical study of social processes. In P. Hedström & R. Swedberg (Eds.), *Social mechanisms: An analytical approach to social theory* (1st ed., pp. 238–266). Cambridge University Press. https://doi.org/10.1017/CBO9780511663901.010

Statistics Finland. (2026). *FIONA remote access system*. https://stat.fi/en/services/services-for-researchers/instructions-for-researchers/using-the-datasets/fiona-remote-access-system

Statistics Sweden. (2026). *MONA - Statistics Sweden's platform for access to microdata*. https://www.scb.se/en/services/ordering-data-and-statistics/microdata/mona--statistics-swedens-platform-for-access-to-microdata/

Stauber, B., Walther, A., & Settersten, R. A. (Eds.). (2022). *Doing Transitions in the Life Course: Processes and Practices* (Vol. 16). Springer International Publishing. https://doi.org/10.1007/978-3-031-13512-5

Tickamyer, A. R., & Duncan, C. M. (1990). Poverty and Opportunity Structure in Rural America. *Annual Review of Sociology*, *16*, 67–86.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *49*, 433–460.

van de Rijt, A., Bernardi, F., Foley, W., & Sage, L. (2026). Luck and Predictability in the Life Course. *Annual Review of Sociology*, *52*(1), 310–339. https://doi.org/10.1146/annurev-soc-031324-115112

Van Der Wel, K. A., Östergren, O., Lundberg, O., Korhonen, K., Martikainen, P., Andersen, A.-M. N., & Urhoj, S. K. (2019). A gold mine, but still no Klondike: Nordic register data in health inequalities research. *Scandinavian Journal of Public Health*, *47*(6), 618–630. https://doi.org/10.1177/1403494819858046

Verhagen, M. D. (2022). A Pragmatist's Guide to Using Prediction in the Social Sciences. *Socius*, *8*, 1–17. https://doi.org/10.1177/23780231221081702

Verhagen, M. D., Stroebl, B., Liu, T., Liu, L. T., & Salganik, M. J. (2025). *The Book of Life approach: Enabling richness and scale for life course research* (arXiv:2507.03027). arXiv. https://doi.org/10.48550/arXiv.2507.03027

World Bank. (2026). *World Development Report 2026: The Promise of Artificial Intelligence*. World Bank. https://doi.org/doi:10.1596/978-1-4648-2331-2

Yan, J., & Rahal, C. (2025). On the Unknowable Limits to Prediction. *Nature Computational Science*, *5*, 188–190. https://doi.org/10.1038/s43588-025-00776-y