UNIVERSITY OF TARTU

Faculty of Social Sciences

Johan Skytte Institute of Political Science

Peeter Leets

AUGMENTING PUBLIC SECTOR DATA-DRIVEN DECISION SUPPORT SYSTEMS WITH EXPERT KNOWLEDGE: CASE OF OTT

Master's Thesis

Supervisors: Mihkel Solvak, PhD

Andres Võrk, MA

Tartu 2022

Acknowledgements

I could not have undertaken this project without my supervisors and dear colleagues, Mihkel Solvak, PhD and Andres Võrk, MA, to whom I would like to extend my heartfelt gratitude. I would like to thank the people from Eesti Töötukassa who were kind enough to help me aquire the required data and expert knowledge for this project. I would also like to thank Taavi Unt, MSc and Bogdan Romanov, MA, for their invaluable insights and guidance; my program managers, Robert Krimmer, PhD, PhD and Liisa Talving, PhD, as well as the defence committee. Last but certainly not least, special thanks go to my significant other, Anna-Lisa, and my friends and family for supporting me throughout this lengthy but engaging journey.

Authorship Declaration

I have prepared this thesis independently. All the views of other authors, as well as data from literary sources and elsewhere, have been cited.

Word count of the thesis: 22,222

Peeter Leets

May 15, 2022

Non-exclusive licence to reproduce thesis and make thesis public

I, Peeter Leets (personal identification code: 39803286022) herewith grant the University of Tartu a free permit (non-exclusive licence) to

1. reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, until the expiry of the term of copyright, my thesis "AUGMENTING PUBLIC SECTOR DATA-DRIVEN DECISION SUPPORT SYSTEMS WITH EXPERT KNOWLEDGE: CASE OF OTT", supervised by Mihkel Solvak, PhD, and Andres Võrk, MA.

3. I am aware of the fact that the author retains the rights specified in p. 1.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Done at Tartu on May 15, 2022 _____ (signature)

Abstract:

Public sector data-driven decision support systems are uniquely challenging to design due to the ramifications they have on the societal level. Accountability and ethical considerations require these systems to arrive at an equilibirium between accuracy and interpretability amid various implementation and data constraints. While these systems need to contribute to legitimate governance through reasoned and explainable decision-making, they also need to accurately model the policy outcomes they were designed to support. Inopportunely, inductive data-driven systems struggle to solve problems that rely on heuristic input. In this thesis, a particular knowledge engineering technique was adopted to augment a public sector Machine Learning decision support tool with domain expert knowledge. The case in question is OTT – a job-seeker profiling tool used by the Estonian Unemployment Insurance Fund to predict the long-term unemployment risks of their clients. Upon augmenting it with knowledge from caseworkers and data scientists associated with the project, some evidence was found that accounting for expert knowledge in probabilistic data-driven models can lead to a model that performs better on new out-of-sample data and is more in line with underlying domain rules. This yields important implications on the future of Machine Learning in the public sector as it opens up new potential use cases in avenues where 1) labelled training data is hard to come by, 2) a more generalizable model is preferred due to frequent changes in the surrounding context, 3) a model has to perfectly mimic domain logic for interpretability and explainability reasons.

Keywords:

Public sector, Machine Learning, domain knowledge, knowledge engineering, decision support systems, expert systems, job-seeker profiling

TABLE OF CONTENTS

1 INTRODUCTION

Automation is the future. Information technologies are increasingly being harnessed for productivity and performance gains in various organizational processes, from high-level management to procedural street-level tasks. Intelligent data-driven solutions have transformed business models across many application domains from medicine to marketing to education (Dwivedi et al. 2021: 2). While automated systems have proved to excel at tasks that are repetitive or involve making inductive generalizations from quantifiable data, they struggle to solve complex, multi-dimensional problems. Machines are known to be inherently probabilistic and lack cognitive judgment invaluable for making reasoned decisions (ibid.: 6). This can render automated systems unsuitable for use cases where significant weight falls on human judgment.

This limitation is holding back innovation in numerous application domains where performance and workflow optimization benefits of automation would otherwise be welcomed. The public sector is one such area where the use of automation systems is under scrutiny due to the ramifications they have on administrative decisions and the inner workings of state agencies (ibid.: 26). Literature on the use of Artificial Intelligence and Machine Learning (hereinafter referred to as AI and ML, respectively) in the public sphere calls for the need to ensure that these systems are fully consistent with the processes they are integrated into. A slew of publications have denoted the importance of human knowledge and value judgment in public sector automated systems[1]. Public agencies have been notably vigilant regarding the adoption of automation technologies. First, the implementation process of these technologies can often be an uphill climb because system developers and procurers struggle to find a common language. While automated systems integrated into public processes must embed appropriate policy goals, miscommunication between technical developers and public procurers often makes this requirement hard to fulfill. Second, as government agencies hold a lot of responsibility and accountability before the public, machines involved in these processes have to produce fully explainable and actionable output in order to contribute to legitimate administrative decisions. Third, as public processes have the power to govern human lives, deployment of machines raises ethical conundrums in this domain.

[1] See: Mulligan and Bamberger (2019); Wirtz, Weyerer, and Geyer (2019); Dwivedi et al. (2021: 28)

A special branch in AI and ML literature known as *knowledge engineering* is devoted to augmenting data-driven systems with human heuristics to address these issues. Although a notoriously complex undertaking, some approaches have been proposed for incorporating human judgment in various stages of model development – from data preparation to parameter weighting (Yu et al. 2007: 17–21). Literature in this field is encouraging researchers to explore different methods of accounting for heuristic knowledge in data-driven models across various application domains[2]. This thesis adopts Gennatas et al. (2020) Expert-Augmented Machine Learning method for public sector use cases and seeks to answer the question – **what are the benefits of augmenting public sector data-driven systems with domain expert knowledge?** This method seeks to exploit qualitative knowledge to identify and penalize system components that are not in line with the logic and business rules of its application domain. Special focus is on the public sector subfield of social policy, where high associated costs, the requirement for evidence-based policy-making, and the weight of human judgment in decision-making add further relevance to the aforementioned issues with public sector automation. The subject for analysis is OTT, a job-seeker profiling tool used by the Estonian Unemployment Insurance Fund to estimate long-term unemployment risks for their clients. Expert knowledge is elicited from a coalition of system developers and unemployment office caseworkers who use the profiling tool in their day-to-day work. Based on the results of prior similar studies, it is expected that the integration of expert knowledge will help to 1) identify confounding model artifacts not in line with labor market domain logic, 2) increase the predictive accuracy of the model, and 3) achieve a more cost-effective model that is able to learn with less training data.

The thesis is structured as follows: chapter two presents a comprehensive overview of ML-based decision support systems, including associated challenges and assumptions relevant to systems designed specifically for public sector use cases. After introducing the case of OTT and providing theoretical justification to how domain knowledge integration can enhance it, research questions and hypotheses are formulated in the end of this chapter. Chapter three provides a detailed overview of the research methodology and data sources. Chapter four is dedicated to analysis and interpretation of the results, while chapter five summarizes key findings and discusses study limitations and potential future research avenues.

[2]See: Sinha and Zhao (2008: 298); Deng et al. (2020: 20); Schapire et al. 2002: 545

2 RESEARCH RELEVANCE AND PREMISES

Present-day society is surrounded by technological artifacts. New inventions are continuously being integrated with our lives to make our day-to-day tasks more efficient and convenient. Technological systems are not only designed and constructed by the society, but they also innately shape the society that employs them (Hughes et al. 2012: 51). Scholars that have studied the societal impact of new technologies have recognized that ideally, technologies should not only aspire to solve issues of efficiency, safety, reliability and ease of use but should also promote and adhere to the social, moral, cultural, and political values of the domain they are integrated into (Flanagan, Howe, and Nissenbaum 2008: 322). Machines and technologies should be judged according to how they convey power and authority in the society, instead of focusing solely on their material outputs like potential productivity gains or environmental impact (Winner 1980: 121).

In his essay on political qualities in new technologies, Langdon Winner (ibid.: 123) considers a vivid historical example where the design of a specific technology contains political properties and becomes a means of settling a particular issue in society. He mentions that for a visitor of the Big Apple, it can seem a little strange how highway overpasses on Long Island (a very affluent neighborhood) are hanging suspiciously low, leaving very little clearance for vehicles from driving under (ibid.). It turns out, as products of their era, they were deliberately designed with low clearance to discourage public transport vehicles to drive on these roads (ibid.). That meant the lower-to-middle class people who relied on public transportation to get about the city (which just happened to contain most of the Afro-American community in New York at the time) were kept out of the district because most buses simply could not physically access it (ibid.: 124).

This would be an example of a situation where a technology *intentionally* embodies political qualities and shapes its surrounding social context by design. It is, however, the *unintended* social impact that is even more difficult to foresee and address in the process of designing new technologies. Acknowledging and mediating unintended and undesired consequences has been of high interest to scholars studying the social role of technologies (Waelbers 2011: 2–3). Waelbers (2011: 3) echoes a famous example of the early energy-saving lightbulbs that were invented to save energy but ended up increasing overall energy consumption because they were cheap to use in places that would not have been lit up in the first place. The same negative effect can occur with technologies that, instead of directly tackling a practical problem, mediate the

decision-making process leading to a potential solution to that problem. As put by Vriens and Achterbergh (2015: 316), responsible decision-making entails solving a particular issue at hand while "minimizing morally relevant side effects". In order to ensure responsible use of a decision support technology, special attention should be cast on which practices and stakeholders will ultimately be affected and whether this technological fix is compatible with associated values and social conditions (Waelbers 2011: 94; Vriens and Achterbergh 2015: 328).

These admonitions are not unwarranted – there have been real-life cases where complex technological agents with too much decision-making discretion have brought about negative social impacts. An infamous case of algorithm-induced bias in a high-profile decision-support technology is the story of the COMPAS risk assessment system for the United States judicial system. In 2016, a team of investigative journalists alleged that the learning-based system that is used to estimate the recidivism risk of criminal defendants is biased against Afro-Americans because the algorithm was found to misclassify one as high-risk more likely than for other groups (Dressel and Farid 2018: 1). This incited a debate among scholars and data scientists over the issue of algorithmic discrimination, with some arguing that the system is actually not biased because the base rates for recidivism differed in the first place (Dieterich, Mendoza, and Brennan 2016: 20–21) (Kleinberg, Mullainathan, and Raghavan 2016: 17). Regardless of the verdict, it is clear that unwanted societal consequences are not to be overlooked when designing technologies that can potentially steer the course of human lives. As I will explain in the following chapters, when and how these concerns should be addressed ultimately rests on our judgments about the use case of a particular technology in question.

2.1 Automation in public processes

In this rapidly developing digital age, governments are more eager to adopt new technologies than ever. Public agencies are undergoing a paradigmatic shift in how they operate, with window clerks, caseworkers, adjudicators and other street-level bureaucrats being replaced by automated expert systems and web sites (Bovens and Zouridis 2002: 175). Information Technology has proven to contribute to public administration by making governments perform more effectively across the board, from internal management to public service provision (Danziger and Andersen 2002: 617). IT-based applications that have generated considerable productivity and efficiency gains range from backroom operations such as digital data-sharing infrastructures and Big Data Analytics (Höchtl, Parycek, and Schöllhammer 2016: 149) to more specialized applications

such as geographic information systems, and public health services (Danziger and Andersen 2002: 593, 605). Over the past decades, the entire novel field of e-government has emerged to cover a variety of use cases and design practices of IT-based solutions specific to the public sphere (Yildiz 2007: 650). Despite the recent surge of associated investments and applications the world over, public sector automation remains a novel subject (Wirtz, Weyerer, and Geyer 2019: 597).

Various definitions of data-driven systems have been brought forward in literature, of which some can be a bit misleading, yet ultimately refer to the same basic concept. For instance, it seems like the "Artificial Intelligence" label gets thrown around a lot in literature to glorify even the least complex of regression models. For the sake of simplicity, I prefer to consider any kind of data-driven applications under the overarching umbrella term of automated systems in this subchapter, dedicated to the challenges associated with these systems in the public sector. Wirtz, Weyerer, and Geyer (2019: 601) identified four major challenges with automated systems that are amplified in the public domain that the author has further synthesized into two overarching themes – implementation barriers and the social aspect – that help to explain the rationale of this thesis.

Implementation barriers

The first implementation barrier with public sector automation concerns data privacy. Generally, the more data is collected, the better and more tailored the resulting service (Bekkers and Zouridis 1999: 190). While the collection and use of personal data for automated services are commonly regulated regardless of the use case or domain of operation, the scope of services renders proper data collection especially troublesome in the public sector. Private companies such as social media websites or banks can freely process their customers' data because they have generally given a consent of some sort. People have a choice whether or not to become their clients and share their personal data. Services and applications in the public domain, however, often assume the availability of data of the entire population because anyone from *the public* is by definition a beneficiary of public services. Take a hypothetical assessment tool that predicts the risk of some person becoming unemployed in, say, the next month in order to preemptively detect potential recipients for a limited amount of unemployment benefits. While it is not a problem to legally gather and process data of people who have registered with the unemployment office in the past, these are not the only people who can theoretically become

unemployed and eligible for benefits in the future. For this application to produce an output that accurately represents the labor market status of a person *relative to the rest of the society*, time-series data from every single citizen, employed or unemployed, would be necessary. At the same time, it would be unthinkable to lock essential public services behind an extensive personal data processing consent as it would be associated with government surveillance the fearsome "Big Brother" state notion (Bekkers and Zouridis 1999: 190).

The second acute implementation challenge regards communication barriers between institutions and system developers. Despite public policies commonly affecting a wide range of stakeholders, it ultimately falls to a small coalition of procurers and system developers to design a system in accordance with the contents of a particular policy (ibid.: 192). As denoted by Bekkers and Zouridis (1999: 192), public sector information systems often end up reflecting what this coalition perceives as relevant. It is the responsibility of public servants and system engineers – two groups of people with inherently different perspectives and knowledge (Bailey and Barley 2020: 4) – to build a model with appropriate business rules, variables, and parameters to serve a functional policy that aims to shape the society in a particular way. It is no secret that information systems favour machine-readable quantitative data over soft, qualitative data (Bekkers and Zouridis 1999: 192). As a result, public organizations are employing machine learning applications that rely on data patterns and inductive reasoning (Mulligan and Bamberger 2019: 778) perhaps more than they should in this case. While proven to offer numerous benefits from prediction accuracy and workflow optimization standpoints, these systems tend to displace human policymakers' knowledge-based judgment with fundamentally probabilistic and sometimes confounding inference that is associated with machines in general. (ibid.).

Recent scholarly attention has converged on the concerning reality that public agencies procuring and employing machine learning systems have often little to no input nor sufficient knowledge of the working principles and capabilities of these systems (ibid.). Procurers have no idea about how different machine learning algorithms predict their outcomes, what variables are used, or even what kind of data are these predictions ultimately based on (ibid.: 778, 801). The reality is that it is not public officials' responsibility to dissect and understand complex machine learning systems, neither can they be expected to have the necessary time and specialist knowledge to do so. However, as illustrated by the COMPAS case, there are legitimate concerns that designing and (mis)using efficiency-oriented machine learning tools in the public domain with little regard to the surrounding context may have a detrimental impact on the society.

Finally, financial feasibility and optimal budgeting is another critical point of concern (Wirtz, Weyerer, and Geyer 2019: 602) as automation projects in the public sector are ultimately publicly funded. The resulting system has to be financially accountable to the public in that the expected benefits outweigh the total associated development and maintenance costs.

Accountability and ethical considerations

The other major challenge regards the social role of automation technologies. Innovation in the public sector can often be an uphill climb because of what is ultimately at stake. Introducing an automated application in the public sector requires a meticulously optimized and specialized strategic approach as a myriad of potential issues can potentially shackle the project in development hell. While the quality of the application and data it utilizes are essential for any IT-based solution, those employed in the public sector must be especially flawless due to their societal ramifications. The data these systems rely on must be accurate, unbiased and relevant, and it needs to be collected and stored properly (ibid.; Mulligan and Bamberger 2019: 796–797). The choice and operationalization of sensitive variables such as gender, ethnicity and race is another particularly important policy decision, as posited by Mulligan and Bamberger (2019: 796). Inadequately designed applications can result in biased policy outcomes such as in case of the above-mentioned COMPAS recidivism prevention tool, where it essentially comes down to the quality of the data and model specification to determine which defendant is detained and who walks free.

A growing body of literature in the fields of human-AI interaction and Machine Learning is putting a strong emphasis on developing explainable algorithmic systems (Amershi et al. 2019; Dwivedi et al. 2021: 7–8) that "provide actionable recourse for the individuals whom they are making [...] decisions about or for" (Troya et al. 2018: 4). The introduction of machines in the governance chain challenges a central principle in public administration – that public actors are to be held accountable for their actions and policies (Bekkers and Zouridis 1999: 191). Opaque and purely inductive statistical systems cannot contribute to legitimate governance as they neglect input from policy-makers and relevant stakeholders (Mulligan and Bamberger 2019: 801). The legitimacy of the state is hinged on the premise that administrative decisions are not taken arbitrarily but through an reasoned deliberation process involving affected stakeholders (ibid.: 804). Applications that have the power to govern human lives (or at least guide the decision-making process) must be entirely explicit, transparent, and fair (Wirtz, Weyerer, and

Geyer 2019: 603). By virtue of minimizing human bias, data-driven systems can often actually improve upon decision fairness (Dwivedi et al. 2021: 26). However, whenever the machine does fail (and it eventually does), reduced accountability of public servants evokes the question – who is accountable when "the computer says no" (ibid.)? While different opinions circulate over who is to be pointed at in this case (Wirtz, Weyerer, and Geyer 2019: 603), the accountability requirement certainly puts extra pressure on developers and procurers alike to deploy a system that does not only minimize machine errors but also allows for these errors to be easily detected and corrected to avoid unwanted social consequences.

It is an intensely debated question whether data-driven models should be used to solve problems that rely on cognitive judgment as some attributes intrinsic to humans are yet to be translated effectively (ibid.: 604) (Dwivedi et al. 2021: 2). While the gross oversimplification that good machines can think and act like humans often makes its way into literature and definitions of Artificial Intelligence (Wirtz, Weyerer, and Geyer 2019: 599), in reality they can at best be trained to mimic human decisions by eliciting empirical correlations from past data using complex mathematics. The way the system is able to "think" or "act" then, depends on the quality and objectivity of the data. As indicated in a previous chapter, data-driven systems are far from ideal in that they are not capable of replicating rational human behavior influenced by consciousness, emotions, and common sense (ibid.: 604). Systems learning from bad training data can echo previous human biases and structural issues in future predictions, potentially amplifying an underlying societal problem instead of fixing it (ibid.: 605), thus neglecting the ultimate principle of governments to treat their citizens on fair and equal terms (Dwivedi et al. 2021: 26). The bottom line is that in order to be fully legitimate, public sector automated systems must reintroduce appropriate expert knowledge to justify their influence on administrative decisions (Mulligan and Bamberger 2019: 818).

High cost of error

Whereas these aforementioned issues hinder the adoption of data-driven systems across the public domain, the effectiveness and viability of a particular data-driven tool are ultimately dependent on its intended use case. Take systems based on the Bayesian classifier – a common type of supervised prediction model employed across many different application domains, including the public sector. The Bayesian approach allows for testing whether a given hypothesis is true or false based on available data and selected influencing factors (Carey and Matthews

2017: 198). Regardless of application domain, the standard performance measures for this type of model tend to be based on a simple confusion matrix that reveals how many positive cases were indeed labelled as positive (true positives) and how many negatives were correctly labelled as negative (true negatives) (Carey and Matthews 2017: 200). While aspects mentioned above are important for determining whether a data-driven tool is appropriate in a certain use case, the most pivotal question for every application should be – what is the cost of machine error and to what avail can this be remedied. No model is perfect – classification errors are bound to happen due to model incompleteness, especially in cases where significant weight is put on qualitative data not easily exploitable by machines.

Comparatively high cost of machine error helps to explain further why the adoption of data-driven solutions has been rather slow and conservative in public sector use cases compared to most other application domains. Tasks that rely on qualitative data and complex heuristics are more difficult to automate. Combined with the far-reaching ramifications of public sector processes, high cost of error can often outweigh potential efficiency gains. A personalized advertisement prediction model that wrongly classifies one's gender and displays an ad for a lipstick instead of a beard trimmer is nowhere near as consequential as a decision-support tool that arbitrates the freedom of a real person. A case in point is (once again) the COMPAS recidivism prevention tool where false positives (defendants not prone to recidivism in reality but detained until trial nonetheless) and *especially* false negatives (defendants prone to recidivism but falsely released) can have huge ramifications for the rest of the society.

2.2 Data-driven tools in social and unemployment policy

Now that general challenges with automated data-driven systems in the public sphere have been outlined, I shall continue narrowing the scope of this study by focusing on the specific subdomains of social and unemployment policy. As the overarching goal of social policy is to enhance welfare by redistributing resources among its citizens as efficiently as possible, the estimation and management of various financial, social and institutional risks is of key importance for governments (Baldock, Vickerstaff, and Mitton 2011: 1–2). I will proceed to list three main reasons why the *correct* use of data-driven systems should be under scrutiny in the making of social (and by extension, unemployment) policy.

The scale and cost of social policy

Welfare provision is expensive. In the European Union, total expenditure on social protection amounted to a hefty 22% of GDP in 2020, with the most significant expenses being pensions, sickness and disability payments, family and children related payments, and unemployment benefits (Eurostat 2022). In addition, 2020 marked the highest expenditure on social protection ever in the EU as means to mitigate the consequences of the global COVID-19 pandemic (ibid.). Particularly noteworthy was the increase in unemployment-related expenditures, which rose from € 180 billion to € 298 billion at the EU level in a span of a single year between 2019 and 2020 (ibid.). High relative cost combined with the magnitude of social services means that exploring ways to make service delivery more cost-effective and *efficient* is always at the forefront of government and public priorities.

More evidence based policy-making

Despite social policy creation being ultimately driven by values and politics, evidence-based policy-making is paramount to achieving accountable and justified resource distribution within the society (Baldock, Vickerstaff, and Mitton 2011: 3). Policy-makers and service providers undergo profound research to determine the most cost-effective intervention strategy that can lead to desired policy outcomes (Davies and Nutley 2000: 121). The prominent role of extensive data-based monitoring and impact assessment studies indicates that the necessary infrastructure for more sophisticated data-driven tools already exists in the field of social policy.

Weight of expert judgment

The primary intended outcome of social policy is what Baldock, Vickerstaff, and Mitton (2011: 11, 14) call the "Robin Hood" function – to redistribute resources from the less needy to the more needy. Meanwhile, the importance of evidence-based data-driven policy-making has already been stressed, this kind of redistribution effort puts a lot of emphasis on ethical and moral considerations so that the result would ultimately be perceived as "fair" and "rational" by the society (Baldock, Vickerstaff, and Mitton 2011: 14) (Kemshall 2001: 20). Value judgments from human domain experts must be embedded into the data-driven system for these decisions to be meaningful, reasoned, and reflective of the policy that it is designed to fulfill (Mulligan and Bamberger 2019: 803–804).

Algorithmic job-seeker profiling in unemployment policy

A key area of social policy where data-driven solutions are showing increasingly promising results in practice, is unemployment policy (Desiere and Struyven 2021: 367). The process of job-seeker profiling involves segmenting clients based on their characteristics and statuses in order to determine the group of people who require imminent intervention by the Public Employment Service (PES) (Desiere, Langenbucher, and Struyven 2018: 1). The general aim is to differentiate the people who are likely to find a job on their own from the disadvantaged people who are in need of support and extra incentives to do so (ibid.). Desiere, Langenbucher, and Struyven (2019: 8–9) have presented three distinct approaches for job-seeker profiling that are currently used or developed in OECD countries:

1. **Rule-based profiling** – this widely used approach involves classifying job-seekers into client groups based on individual factors such as age, education or unemployment spell duration (Desiere, Langenbucher, and Struyven 2019: 8) – for example, some employment benefit may be made available for all people over the age of 50 with a certain level of education. While this method is simple to understand and implement, most applications of rule-based profiling need to be combined with caseworker judgment or further assessment tools to make the service provision as personalized as possible (ibid.). For how easy it is to understand, a typical rule-based approach still produces relatively heterogeneous client groups that cannot capture many potentially important personal details about an individual that a human caseworker would be able to.

2. **Caseworker-based profiling** relies on human caseworkers' judgment in classifying job-seekers into priority groups (ibid.). Caseworkers can either be given full discretion in judging clients' risks and needs (as it has previously been done in Estonia), or more typically, their decision-making may be supported by additional analytical tools (ibid.). The obvious benefit of this approach is that the decisions are tailored to each client. The downsides of this approach are cost-ineffectiveness as it requires a lot of human workforce to deliver, and the fact that decision-making validity ultimately hinges on their competence and objectivity.

3. **Statistical (algorithmic) profiling** uses an inductive statistical model to predict output related to clients' labor market status (ibid.: 9). Statistical profiling has the advantage of estimating individual risk scores for each job-seeker as opposed to classifying them

according to pre-defined decision rules (Desiere, Langenbucher, and Struyven 2019: 9). At the same time, it remains much more resource-effective than caseworker-based profiling (ibid.) as it is ultimately a quantitative computational method.

Statistical and algorithmic profiling of job-seekers has recently become a hot subject for applied research, with quite a few independent methods being simultaneously designed and tested. Troya et al. (2018) used machine learning models with different levels of complexity to estimate long-term unemployment risks. They demonstrated the trade-off between model accuracy and interpretability and highlighted the importance of decision explainability in automated job-seeker profiling and public processes in general (Troya et al. 2018). Viljanen and Pahikkala (2020) developed a Markov chain model to estimate risks related to people's unemployment statuses based on individual-level registry data obtained from the employment and business service of Finland. Actual in-use profiling models have been rolled out in different countries as well. Since 2020, the Austrian Public Employment Service (AMS) has used a logistic regression-based profiling system that classifies job-seekers into support categories based on a statistical model of individual labor market prospects (Allhutter et al. 2020: 2; Desiere, Langenbucher, and Struyven 2019: 12). Different variations of ML-based statistical profiling tools have also been in use in Australia, Belgium, Denmark, Ireland, Italy, Latvia, New Zealand, Sweden, the United States (ibid.) and most recently – Estonia.

The Belgian PES is one prominent case where an ML-based statistical profiling tool has been piloted in a real public unemployment service process. The Belgian system is intended to assist employment office caseworkers in deciding which job-seekers to prioritize by estimating the probability of becoming long-term unemployed for each individual (Desiere and Struyven 2021: 371). Under the hood, it operates on a Random Forest model, which is continuously being trained with standard socio-demographic data such as job-seekers age, education, and nationality, and more specific employment-related information, including their previous job spells and participation in training programs (ibid.). The primary use-case of the Belgian system is to determine which job-seekers are most at risk and should therefore be contacted first (ibid.: 372). Thus, the tool does not determine the type of support that the job-seeker receives but merely ensures that the most vulnerable people are addressed as soon as possible (ibid.).

Just like for the COMPAS case mentioned in chapter 1, concerns over algorithmic discrimination have emerged with automated job-seeker profiling. Allhutter et al. (2020) critical reflection on the Austrian profiling tool discusses issues with accountability and transparency

in its working practices. Among their key points is that despite advertised gains in accuracy and efficiency, decision-making responsibility ultimately still lies on human caseworkers to correct the occasional mistake made by the system (Allhutter et al. 2020: 14). By shifting to an algorithmic solution, evaluation of human job-seekers ends up being based on quantifiable data, and fails to take into account cognitively detectable aspects including "soft" skills and motivation (ibid.). Because there is a constant need for model validation, caseworkers are torn between either putting full discretion to the automated prediction system or going back to trusting their own experience and judgment (ibid.). It needs to be pointed out, however, that as opposed to the Belgian system that is used for determining contact priority, the Austrian system was given full discretion to determine what kind of support should be offered to certain people. The Austrian system classifies people into three categories based on their risk score, which instantly dictates the type and intensity of provided support measures (ibid.: 2).

2.3 Machine Learning enabled decision support systems

Many decisions are based on predictions for the future. Companies hire employees based on how productive they predict them to be, banks give out loans based on how likely the client is to pay them back in the future, and investors put their capital in companies that they believe will do well in the future. Some of these real-world problems are relatively easy to predict – for example, when there are not too many factors to be considered and one or two factors are hugely decisive. When one witnesses a completely clear blue sky as far as the eye can see, one would be quite confident that the chance of imminent rainfall is small to none. More complicated prediction problems require a deeper analysis of multiple factors and their possible interactions. Predicting a Formula One race winner requires considering a range of factors – from the skill and motivation levels of the drivers, to changing weather conditions, to different race strategies, to performances of certain cars on certain tracks, and so on. The more factors are included, the more complex the model grows, introducing questions like how much weight should each factor be assigned to predict the outcome? Is the value for some predictor variable dependent upon the value for some other predictor variable(s)? For example, if some drivers drive better in the rain than others, their skill levels would be conditioned upon the weather variable, and this interaction would somehow have to be accounted for in the model for accurate output.

There is a famous saying in the field of statistics, most often attributed to British statistician George E. P. Box (1976: 792) that all models are wrong, but some are nonetheless useful. In other

words, in most cases of statistical modelling, we end up making tentative assumptions about the real world, which we know are incorrect or over-generalized (Box 1976: 792). Yet by treading carefully, researchers may still be able to reach a relatively valid and informative approximation of the situation (ibid.). There is virtually no hope for researchers to ever accurately predict highly dimensional problems such as the stock market or the result of a sporting event (being able to do so would render these activities essentially pointless anyway). Still, even the most daunting problems can theoretically be modelled with a reasonable degree of success as long as the interpreter knows roughly when and where the model tells the true story. Bookies capitalize on experts' knowledge to try and predict the outcomes of sports events accurately enough that it is profitable for them to do so, whereas financial analysts rely on their domain knowledge and market history in predicting the general trends that the market should adhere to in the long run. It is a notoriously difficult problem to predict how a stock will perform in the next week, day, or hour. Predicting its performance over the long term is a lot easier, however, as it allows the analyst to examine long-term patterns in the company's quarterly financial results, the performance of the sector it belongs to, together with various macroeconomic factors.

Advances in the relatively young and novel field of Machine Learning have proven to be helpful in tackling some of those complex, multi-dimensional prediction problems. Machine Learning is a breakthrough in data science that allows analysts to detect complex and highly dimensional patterns in data (Kleinberg, Lakkaraju, et al. 2018: 238). The fundamental goal of machine learning is to make computer systems learn from available data and use this knowledge to improve their decision-making accuracy for future predictions (Jordan and Mitchell 2015: 255). ML methods and applications vary greatly, to the point where the term itself has become somewhat overused. At one end of the spectrum, complex ML methods such as deep neural networks have become integral components in different artificial intelligence (AI) frontiers by facilitating speech recognition, visual detection and natural language processing through learning from training data (ibid.). At the other end, the term can technically be used to glorify simple regression models that link a set of predictors to an outcome variable and are then used to predict new output for previously unseen data.

For the purpose of this thesis, I will focus on a particular branch of ML systems – Supervised Machine Learning. Supervised (also referred to as inductive) ML systems typically estimate a set of complex functions that link a desired output variable Y to a set of various predictor variables X (Molina and Garip 2019: 28). The set of functions (the model) can then be applied to new

(unseen) data to predict the outcome variable Y for that data (Molina and Garip 2019: 28). Yu (2007: 3) describes the inferential process behind inductive machine learning in two relatively straightforward steps:

1. *"A learning system L constructs a hypothesis space H from a set of training examples (or observations) S, for example a set of training examples consisting of input-output pairs $S = \{\langle x_1, c(x_1)\rangle\}$, and the inductive bias B predefined by the learning system and the task T;"*

2. *"the learning system then searches through the hypothesis space to converge to an optimal hypothesis f consistent with training examples and also performing well over other unseen observations."*

In layman's terms, a supervised ML algorithm is given the opportunity to learn or *train* from using human-labelled data – for example, if training data indicates that for ten recorded clear days, it only rained once, then the algorithm predicts that in the future it is likely not going to rain on a clear day. Of course, a problem worthy of ML application usually involves a vector of many potential predictors, which makes it hard for humans to estimate the outcome as reliably and accurately as a trained machine would.

Supervised ML prediction algorithms can only be as reliable as the data they are trained with. It is unlikely that the data set used for training or "teaching" the model covers the whole population (Yu et al. 2007: 4). In other words, the algorithm must often predict the outcome for observations for which there does not exist an ideal match from the training set, meaning the algorithm has to make some sort of a generalization based on other, similar training observations. When the training set is rather small, the model has to generalize more, resulting in lower prediction accuracy. At the same time, if the model is set to account for as much variation in the training data as possible, the model runs a risk of *overfitting* – generalizing too little. Such a model becomes too specific to the underlying data structure of the training set and, as a result, performs poorly on a new unseen set of observations (ibid.). This trade-off is known as inductive bias in the Machine Learning community (also known as the bias-variance trade-off in statistics), and it presents one of the most problematic questions for ML system designers: how to find the optimal hypothesis space for the learning task so that it is large enough to solve the problem at hand, yet small enough to ensure that the solution remains generalizable for future observations (ibid.: 5; Baxter 2000: 149).

Modelling constraints are just the tip of the iceberg, however. As machines and AI-based solutions are gaining an increasingly important role in human lives, the use of ML applications has introduced broader challenges that go beyond mathematics and data science. Different contexts and use cases often require prediction tasks to be approached on a case-by-case basis, as constraints set by available resources, data privacy or the social, legal and political environment that the system is eventually integrated into, tend to vary considerably (Jordan and Mitchell 2015: 255). Certain use cases may require the system to be easily interpretable, explainable, and/or visualizable (ibid.), achieving which is often a surprisingly difficult feat when engineering ML applications. For example, an ML algorithm that predicts the likelihood of getting some disease should ideally not be black-box. While it may be considered valuable to highlight the group of people who are most at risk based on a simple probability score, it would be crucially important to understand *why* their risk is as high as it is and which factors contribute most to each individual score, as this additional information can determine the type of treatment that is ultimately offered to the individual. Increased model interpretability usually comes at a cost, however. It has been established that in AI and ML applications, there is an inherent trade-off between model accuracy and interpretability because the underlying interactions become less comprehensible as the complexity of the model increases (Dwivedi et al. 2021: 8). This trade-off is especially noticeable for modern deep learning systems such as neural networks, which make the task of model explanation notoriously difficult for analysts (ibid.). Ultimately, there is little use for a data-driven system that does not produce fully interpretable, actionable output, that should ultimately be the basis for evidence-based policy intervention.

Today, more and more organizations are trying to leverage the potential of Machine Learning technology to optimize their working and decision-making processes (Edwards, Duan, and Robins 2000: 36). While earlier literature about Artificial Intelligence and technological transformation revolves around AI replacing human workers, more recent accounts recognize the limits of automation and take a more realistic view – AI should ideally enhance and optimize human capability, not replace it (Dwivedi et al. 2021: 4)(Bailey and Barley 2020: 3). AI-enabled expert systems can theoretically be employed at all organizational levels, but historically, expert systems have proven to be most effective at the lowest operational levels, where problems tend to be the most structured and therefore the most predictable (Edwards, Duan, and Robins 2000: 44). Expert systems can also be used to support decision-making at higher organizational levels, however, as problems become more unstructured (such as high-level strategic decisions),

the effectiveness of expert systems generally wanes off due to higher decision uncertainty and complexity (Edwards, Duan, and Robins 2000: 44). The more complicated and qualitative the nature of the problem is, the harder it is to design and exploit a reliable data-driven expert system.

Ultimately, however, there are a number of traits that cannot (yet) be acquired by machines at all, including cognitive skills, critical thinking, creativity, and intuition (Deng et al. 2020: 1; Dwivedi et al. 2021: 6). Errors caused by the overly probabilistic behaviour of machines are relatively common, which is why building explainable and interpretable systems has been a key avenue in AI and ML research (Amershi et al. 2019: 2). One of the most ambitious and highly researched problems in the field of Human-Computer Interaction (HCI) is how to integrate human knowledge and experience into data-driven models. The goal is to get the best of both worlds – efficient and robust decision-making process *and* accurate and explainable decisions. Successfully integrating human cognitive knowledge into machine learning models can yield a number of important benefits, such as fewer data would be required for accurate and reliable predictions, but perhaps most importantly, human-in-the-loop systems tend to be easier to interpret and explain than systems that are purely data-driven (Deng et al. 2020: 1). To maximize value creation and minimize undesired social consequences stemming from the overly probabilistic behavior of AI, it is best to utilize AI technologies in a way that they result in a synergy with human workers, with the former doing the calculations and the predicting and the latter being responsible for analysis and interpretation of the results (Yang et al. 2020: 1; Amershi et al. 2019: 1–2; Dwivedi et al. 2021: 4–7; Duan, Edwards, and Dwivedi 2019: 63).

2.4 Domain knowledge in Machine Learning

Designing AI and ML for specific application domains can be difficult for data engineers because, despite their best intentions to inform themselves of the problem they are working with, they are often left ill-equipped in terms of domain-specific knowledge and experience (Yu et al. 2007: 1–2; Bailey and Barley 2020: 5). At the same time, experts of the respective domain are usually not able to comprehend complex Machine Learning models and their working principles, meaning that it is just as difficult for them to combine data-based information with their qualitative domain knowledge (Yu et al. 2007: 1–2). Matters are further complicated by the fact that, as experts themselves often find it difficult to explain *what* is it exactly that their special "rule of thumb" knowledge entails, it is challenging to acquire it in the first place

(Sinha and Zhao 2008: 287). Studer, Benjamins and Fensel (1998: 163) add that even when experts are capable of articulating their knowledge, some skills and experiences may remain hidden in their subconscious. Human heuristic knowledge is regarded as highly valuable in the context of many prediction problems, and so the question of how to integrate it with data-driven models as conveniently as possible remains elusive. This chapter introduces the concept of domain knowledge (sometimes referred to as expert knowledge – this thesis uses these concepts interchangeably) and covers topical literature that involves integrating it with machine learning applications.

In an effort to synthesize this relatively vague concept into a more tangible and measurable construct, Alexander (1992: 34) defines domain knowledge as "the realm of knowledge that individuals have about a particular field of study". Yu et al. (2007: 9) definition ties it into the context of AI and Expert System design:

"the prior domain knowledge is all of the auxiliary information about the learning task that can be used to guide the learning process, and the information comes from either some other discovery process or domain experts."

It is usually relatively straightforward to consider domain knowledge that can be quantified and represented in an equation-based format. The simplest example of exploiting quantitative domain knowledge in machine learning is explicitly defining some rule that sets constraints on the training process, for example that $Mass = Density \times Volume$ (Deng et al. 2020: 5). In many cases, however, domain knowledge is abstract and non-quantifiable, which makes it difficult to integrate with machine learning frameworks operating exclusively with quantitative data (ibid.: 2). Knowledge based on observation, logical inference, and induction typically requires the qualitative input of people with domain-specific experience and skills (ibid.: 5). At the same time, experts and specialists of one domain operating at different organizational levels may develop different kinds of domain knowledge and experience (Yu et al. 2007: 9). For example, a labor economist and an unemployment office desk worker are largely knowledgeable in the same umbrella domain that is labor economics. However, the former is likely to have better macro-level knowledge of prevailing trends in the labor market, while the latter is more likely to accurately predict employment for individual job-seekers based on direct interaction with them.

Numerous efforts have been made to integrate domain knowledge with Machine Learning – a process known as *knowledge engineering* (Sinha and Zhao 2008: 288). The problem of combining domain knowledge and inductive data-driven learning models has proven to be notoriously

complicated and there is no universally "right" way to do so (Yu et al. 2007: 1). Prior applied studies have adopted highly customized *ad hoc* methods depending on the domain context, availability of data, and perhaps most importantly, the use case of the designed application. It is important to note that the process of knowledge engineering does not entail building a perfect cognitive model that perfectly replicates human heuristics but a model that replicates the problem-solving process as adequately as possible in a given application domain (Studer, Benjamins, and Fensel 1998: 163). As mentioned above, experts may not always be conscious about their skills and experience, much less articulate them in a quantifiable way. These (often crucial) parts of expert knowledge are not directly accessible to system designers, "but have to be built up and structured during the knowlegde-acquisition phase (ibid.)". According to Studer, Benjamins and Fensel (1998: 163), such an approach should be viewed as a **part of the modelling process** itself rather than a separate *ex-post* transfer of knowledge.

Yu et al. (2007) acknowledge four types of methods to account for domain knowledge in machine learning models. The first category includes methods that rely on prior domain knowledge in the earliest phase of model design to select, transform, and prepare the data used for training the model (Yu et al. 2007: 17). This approach is especially handy when the data is noisy, and there are many, possibly redundant and overlapping variables to consider (ibid.). As brought out earlier, an inductive learning algorithm can only be as useful as the data it relies on to make predictions. In specific application fields, domain experts' profound knowledge of underlying processes can be useful for sanitizing the data set, including detecting outliers and redundant or missing observations. In some cases, training data even comes in certain underlying structures, such as in the form of a tree or a network, which has to be detected, validated, and transformed to a suitable format by domain experts (ibid.: 18).

Enhancing the data preparation process using domain knowledge has been hugely prosperous and beneficial in medical studies. One reason for this is that in medicine, laboratory samples usually involve a very marginal amount of clear-cut positive cases, which means that the training set is far from the ideal representation of whatever real-life scenario is being modelled (Mirchevska, Luštrek, and Gams 2014: 163). Additionally, the proportion of positive cases itself may often be incredibly small (in case of some rare disease, for example), which further hinders classification accuracy and gathering a sufficiently representative training set. Another reason is that in medicine, raw data (for example, electronic health records) is often not in a "flattened table" format that is typically suitable for data analysis and learning methods (Lin and

Haug 2006: 489). Lin and Haug (2006) proposed a part metadata, part domain knowledge-based data preparation framework for sanitizing noisy and redundant data for learning algorithms and successfully applied it in the development of a decision support system. Coulet et al. (2008) presented an ontology-based data selection method that takes advantage of underlying assumptions and associations in the data set. The works of Rajagopalan and Isken (2001: 466) and Soibelman and Kim (2002: 47) also proved that domain knowledge can contribute to effective data preparation. Finally, Sinha and Zhao (2008) combined knowledge engineering and data mining and showed that the performance of ML models can be significantly improved through integration with expert knowledge.

The second category of Yu et al. (2007: 19) entails exploiting domain knowledge to construct the hypothesis space for the learning model. Given that the hypothesis space represents the set of all possible functions available for the training data, domain knowledge can be helpful for seeking out the best hypothesis and, therefore, the best function for predicting the desired outcome. This can be done by constructing the initial hypothesis space on the basis of which hypotheses satisfy prior domain rules (Mirchevska, Luštrek, and Gams 2014: 164).

The remaining approaches employ domain knowledge to modify the search for the optimal hypothesis (function) itself that is ultimately extracted from the hypothesis space. Purely inductive learning algorithms solve their prediction problems by finding the best objective function that minimizes prediction error, without direct regard to underlying theoretical or logical assumptions (Yu et al. 2007: 21). Domain knowledge can be used to alter or augment the search for the optimal function by introducing additional regularizers or constraints to the objective function itself (ibid.). In practice, this can be done by either setting learning constraints or, more often, weighting the training observations' influence based on some external domain rules (Mirchevska, Luštrek, and Gams 2014: 164). Cao and Tay (2003) forecasted financial time series using Support Vector Machines (a complex non-parametric learning method) augmented by adaptive regularization parameters. These parameters were set to place more weights on recent training observations because, in this domain, recent data is considered more influential for predicting future trends (Cao and Tay 2003: 1513, 1517). Their results showed that model augmented by domain knowledge are able to achieve better generalization performance while using less data (ibid.: 1517). Schapire et al. (2002) similarly integrated prior knowledge with a tree-based ML algorithm by weighting model parameters according to pseudo-examples estimated by human experts, which also proved to improve performance for fewer data.

Proving most relevant to the thesis at hand, Gennatas et al. (2020) introduced another parameter weighting method to enhance probabilistic machine learning models with expert knowledge and tested it by predicting mortality risk from physiologic data. They used a prediction rule ensemble model aptly named RuleFit to extract a set of easily interpretable decision rules from the data and asked experts to assess the risk of subpopulations defined by each rule (Gennatas et al. 2020: 4575). They then ranked these rules based on the extent of expert and model disagreements in assessing subpopulations' risks (delta rank measurement) (ibid.: 4573). Finally, they penalized decision rules with the highest delta ranking by excluding them from the final model (ibid.: 4572). This resulted in those rules having more predictive influence, where experts and the empirical model agreed with each other.

Gennatas et al. (2020) method proved to be effective for several reasons. First, this approach was able to discover what they call *hidden confounders* in training data (ibid.: 4575). In statistics, confounders (or confounding variables) commonly represent unobserved factors that create bias in causal effect estimation (Greenland, Pearl, and Robins 1999: 29). Gennatas et al. (2020: 4571) bring the example of a study where a learning algorithm was configured to estimate the probability of death from pneumonia. The resulting algorithm predicted a lower mortality risk for asthmatic patients than for non-asthmatic patients, a correlation that is certainly misleading but nonetheless based on actual data (Gennatas et al. 2020: 4571). The *hidden confounder* in this case is that asthmatic pneumonia patients are treated much more aggressively (rightly so), resulting in a lower mortality rate overall (ibid.). A learning algorithm is not able to detect such an artifact when the level or even the presence of treatment is absent from the model. By letting experts virtually compete with the empirical model in estimating subpopulation risks, Gennatas et al. (2020) method effectively helps to discover cases where some important predictor is clearly missing from the model. Alternatively, consistent disagreements in regard to the risk of some specific subpopulation can hint at limitations of experts' knowledge instead of hidden model bias (ibid.: 4575). Their approach also resulted in better generalization to new data while requiring less data to achieve comparable prediction accuracy to base models (ibid.). Finally, they found that expert knowledge considerably improved model performance when tested on a population whose variables changed over time or were collected at a later time (ibid.).

It is important to note that incorrect domain knowledge, or knowledge that only partially covers the domain, may yield unsatisfactory results and, in some cases, even harm the predictive performance of the model (Yu et al. 2007: 2). Therefore it is necessary to arrive at an optimal solution where domain knowledge can mediate hidden confounders in training data and improve model generalizability while the benefits of data-driven predicting remain intact.

To sum up the preceding subchapters, three main areas of improvement can be listed when the theoretical benefits of integrating domain knowledge with statistical tools (hereinafter referred to as expert-augmentation) are put into the context of automation in the public domain:

1. **Expert-augmentation can make automated tools contribute to accountable and legitimate decision-making**. As explained in subchapter 2.2, the use of automated tools in public processes automatically becomes more justified simply by virtue of containing the judgment of human decision-makers. While the legitimacy of an ML application is not to be ultimately judged in this thesis, this benefit should be reflected in improvements in the predictive performance of the model. As prior studies have recorded notable accuracy gains on expert-augmented models compared to purely statistical ones, improved accuracy can be taken as a sign that expert knowledge helps to bring the model in line with domain rules and policy goals.

2. **Expert-augmentation can make automated tools operate more efficiently.** Financial feasibility was brought out as an important requirement in public sector automation projects. As shown by prior studies listed above, expert-augmented models are often able to achieve comparable or better predictive performance on less training data. Reducing the required minimum amount of training observations would make automated systems more cost-effective and expand applications to use cases where pre-labelled data is hard to come by.

3. **Expert-augmentation can make automated tools more explainable.** Machine Learning models are notoriously difficult to interpret. The behaviour of automated tools in the public sector, however, needs to be fully comprehensible in order to support administrative decision-making. Machine Learning systems infused with qualitative domain knowledge should theoretically produce more interpretable and actionable output. For example, by penalizing those system components that produce confounding output in the vein of Gennatas et al. (2020), a simpler, more interpretable model is constructed.

2.5 Case of OTT

This subchapter introduces OTT, the subject for this applied study. OTT[3] is an ML-enabled decision support system for profiling job-seekers jointly developed by the Estonian Unemployment Insurance Fund (*Eesti Töötukassa*, hereinafter referred to as UIF) and a team of researchers, labor economists and data scientists from the University of Tartu. The case of OTT was chosen for its relevance in regard to the issues mentioned in the last chapter. As with other public sector machine learning tools, its predictive accuracy is under scrutiny as its output can influence administrative decisions. OTT operates on a rather complex ML method (introduced in subchapter 2.5.1), making it challenging to ensure that its technical system accurately conveys important unemployment policy and domain objectives. This, however, is a crucial requirement, as the policy objective that OTT seeks to support relies heavily on human judgment, as explained in subchapter 2.5.2. Let it be mentioned that, while in theory, most public sector machine learning tools match these criteria, OTT was chosen in this case to support the feasibility of this project from two aspects: 1) the author had access to invaluable first-hand knowledge about the development and inner workings of OTT as a part of the development team, and 2) the quality and sheer extent of expert input from *Töötukassa* caseworkers (explained in chapter 3.1.2) was greatly enhanced by the author's ability to communicate in their native language.

As the subject matter was developed and piloted rather recently, research on the subject matter is lacking. Consequently, the bulk of *a priori* knowledge about the use case and inner workings of OTT originates directly from system developers who incidentally supervised this thesis. In terms of its working principles, goals, and use cases, OTT bears a lot of resemblance to the Belgian system. After a brief piloting period in the first half of 2020, OTT was officially integrated into unemployment office caseworkers' workflow in October 2021. The next section provides an overview of the technical aspects of the model, followed by an explanation of its use case – how it is *intended* to be used and interpreted by caseworkers.

2.5.1 Technical background

Under the hood, OTT is a Random Forest based ML model that, at the time of writing, includes a total of 45 variables that are proven to offer predictive value in estimating job-seekers' future labor market prospects. The underlying model has gone through various updates and changes regarding its inputs and outputs throughout the period that OTT has been in active use (including

[3] ***Ot****susustus****t****ugi*** – Estonian for "decision support"

Table 1. Categorization of input variables in OTT.

Category	**Examples of variables**
Socioeconomic characteristics	Gender, age, citizenship, county of residence
Motivation to look for/accept job	Information about last employment spell(s): total number, length, time since, field, reason for ending, etc., way of registering as unemployed (online or at the bureau), assigned unemployment allowance and UIB, frequency of receiving salary within the last 2 years
Job readiness	Level, field and type of education, work capacity, Estonian language skills, computer skills, has e-mail account, has driver's license, belongs to risk groups (released from prison, caretaker, etc.), is board member, participation in training programs
Opportunities	Amount of job requests and clients registering as unemployed at the same time, job vacancy rate for suitable positions, proportion of job-seekers exiting unemployment status in the last 30 days, amount of unique employers in the last 3 years

the pilot project). Initially OTT launched with as many as 63 predictor variables and seven different outputs, which over time has been reduced to 45 predictors and one concise output. At this time, the model output value is defined as **the probability of exiting unemployment status within the next 180 days after registering as unemployed**. Input variables range from basic socio-demographic factors such as age, education, and region, to more specific labor market variables such as information about previous employment spells, job vacancies and claimed unemployment benefits. Examples of predictor variables consistent with Desiere, Langenbucher, and Struyven (2019: 13) profiling requirements are presented in table 1, an exhaustive list with variable types and verbose explanations can be found in appendix 1.

As mentioned, OTT is based on Random Forest – a popular ML method that has been adapted in job-seeker profiling before, most notably in the Belgian application. The following section is dedicated to briefly examining the working principles of the Random Forest model. I shall start by explaining the cornerstone of decision forest models in general – a single decision tree. Building a decision tree is a relatively straightforward process of splitting observations in two step-by-step based on the characteristic (variable) that offers the greatest homogeneity as a result (Speybroeck 2012: 243). A simple example of a decision tree can be seen in figure 1 that predicts the average fuel consumption of a hypothetical set of cars based on three factors: type

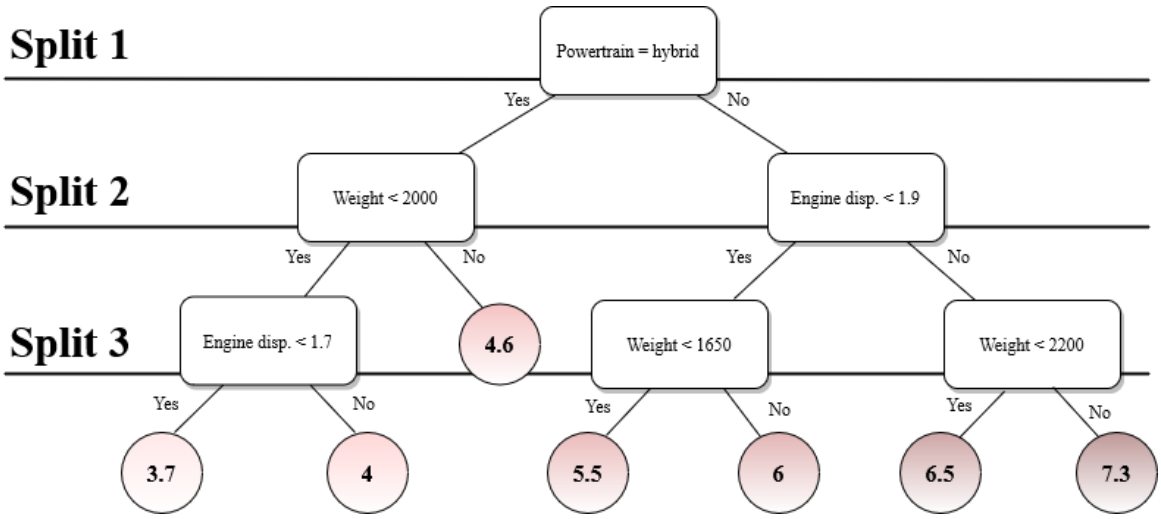

Figure 1. Example of predicting fuel consumption of different cars using a decision tree.

of powertrain, engine displacement, and curb weight. It can be seen that there are a maximum of three steps for splitting the cars into subgroups. First, the biggest factor that appears to determine a car's fuel efficiency is the type of its powertrain. Then, depending on whether the powertrain is hybrid or not, the next splitting variable would be either weight for hybrids or engine displacement for non-hybrids. In every split (internal node), the algorithm chooses a splitting variable so that the resulting two subgroups would differ from each other as much as possible. For regression trees, the resulting mean value for each terminal node is the prediction for that particular group of observations.

On its own, a single decision tree does not yield very good predictive accuracy on new data because the tree splits closely follow the properties of the particular training set that was used to build the tree. Random Forest, as its name suggests, combines many decision trees into an ensemble to increase the predictive accuracy of a single tree in order to improve predictive accuracy while simultaneously reducing overfitting to training data (Montgomery and Olivella 2018: 734). The randomness that is referred to in the name of Random Forest comes from the way that the model handles variable selection at each split. As opposed to a regular decision tree, at each split node, the RF model is allowed to split observations only based on a randomly picked subset of all predictor variables (Breiman 2001: 11). For example, in figure 1, instead of splitting observations based on type of powertrain at step 1, a RF tree may instead be allowed to only split them either by weight or engine displacement, if by chance, powertrain was not allocated into the subset of predictor variables for that particular split. By constraining the selection of possible splitting variables for each tree and node, the individual trees become

less correlated, which in turn makes the model more flexible and generalizable for new data (Montgomery and Olivella 2018: 734).

An important nuance regarding how Random Forest is implemented in OTT needs to be mentioned. While the model is essentially set up to solve a classification problem (predicting a binary representation of whether a person exits unemployment in the next 180 days), it instead outputs the *probability* of the outcome event *not* happening, interpretable as the risk of not resuming work in 180 days. The need for probability scores, as opposed to simple yes-no classes, is explained in the next subchapter that describes the intended use case of OTT and similar decision support tools used in welfare provision.

2.5.2 Use case and potential areas of improvement

As OTT has been in active use for a relatively short period of time, it remains to be studied how exactly caseworkers integrate OTT with their workflow in practice. However, as the fundamental working logic of OTT closely follows that of the Belgian system, the assumption may be made that the use case is, for the most part, similar as well. The Estonian caseworkers are *intended* to use the tool primarily for creating a virtual priority order in their client portfolio – job-seekers who are most at risk of becoming long term unemployed are offered services more intensively. Therefore, the myriad of ethical and fairness concerns that emerged with COMPAS or even the Austrian profiling system are avoided, as will be explained.

Just like the COMPAS recidivism prevention case in the US, the Belgian and Austrian job-seeker profiling systems have received criticism for potentially discriminating and stigmatizing certain types of people, as well as amplifying existing underlying inequalities in the labor market (Allhutter et al. 2020: 2; Desiere and Struyven 2021). Desiere and Struyven (2021: 380) explain why, for the Belgian job-seeker profiling tool (and by extension, OTT), these concerns do not really matter. The key argument boils down to the use case of the system – what are these systems used for, and what kind of consequences do the decisions made by the system have. As Desiere and Struyven (2021: 380) astutely point out, concerns over labor market discrimination are not as severe as often portrayed because the services that these algorithms help to distribute are ultimately considered helpful and benevolent. The use case of such decision support tools is not to decide whether an individual should be punished but instead whether they should be offered help. Indeed, it raises questions when a COMPAS-like system systematically predicts higher recidivism rates for Afro-Americans as simply *being* Afro-American should

not increase one's chances of being detained. For that reason, it makes sense to avoid using so-called protected variables such as ethnicity and race in statistical tools because their presence can reinforce historical biases and existing inequalities encoded in training data (Allhutter et al. 2020: 7). When we consider OTT or other decision support tools in social policy, excluding sensitive variables as a conscious design choice can actually produce the opposite effect. If in some hypothetical society, a racial minority has historically been less likely to find a job on their own compared to the rest of the population, then it is important that the algorithm is able to detect that artifact so that the employment service can remedy this underlying social inequality. Excluding race from the model would ultimately enforce inequalities instead, as the underlying relationship would be largely undetected by the algorithm and, in turn, neglected by the employment service and their support mechanisms. Job-seeker profiling tools (and other automated tools in social policy, for that matter) circumvent this caveat with their specific use case – what they are ultimately being used for. While a recidivism prevention tool that systematically points at individuals from a certain disadvantaged group can (and arguably must) be construed as unfair, an equivalent tool in social policy *should* point at these groups to ensure that government aid reaches those who need it most. The goal of OTT and other job-seeker profiling tools is to find people that are disadvantaged or discriminated on the labor market. If it emerges that a certain disadvantaged group forms strongly based on some sensitive variable, then so be it – in the end, these people will ultimately benefit from this discovery.

That is not to say that there is no room for improvement for OTT and other automated job-seeker profiling tools, however. While the rather cautious use case of OTT (it is merely supposed to rank clients based on the need for intervention) generally renders aforementioned concerns over its algorithmic fairness moot, one point of criticism towards similar applications in this domain remains acute. Briefly touched upon in previous chapters, there are certain factors that are difficult to accurately measure and/or include in quantitative models, yet are known to have a considerable influence on the outcome – in this case, one's likelihood of finding a job. Allhutter et al. (2020: 11) mention the inability of the Austrian system to capture job-seekers' soft skills, engagement and motivation in a quantifiable way – features which are traditionally considered by caseworkers through a more customer-oriented profiling approach. At this time, the Belgian system does not consider soft skills, personal attitudes and job search strategies either (Desiere and Struyven 2021: 372). Desiere, Langenbucher, and Struyven (2019: 14) also denote that while labor economics literature has emphasized the importance of behavioural factors in

determining employability, they are notoriously difficult to directly measure and operationalize in statistical models. Furthermore, there is some evidence that including behavioural "soft" variables in profiling models does not offer a substantial increase to most profiling models because usually, the model already includes other "hard" variables that are strongly correlated with them (Desiere, Langenbucher, and Struyven 2019: 14).

Another important caveat with automated job-seeker profilers (including OTT) has to do with their (in)ability to capture abrupt shifts in the surrounding economic context. Automated data-driven decision support tools can only make predictions based on data that has been collected in the past. Unfortunately, this data might not always represent the future as accurately as desired, especially in this particular application domain. Job-seeker profiling tools need to be constantly updated with the latest data that best models the most recent trends and changes in the continuously changing labor markets (ibid.: 23). Certain characteristics of job-seekers that in the past were strong indicators of a person finding a job quickly might not be good predictors today if, for example, that line of work is not in high demand anymore (ibid.: 16). It is not technically difficult to recalibrate profiling models with new data *per se* as it is done for the Belgian system, for instance (ibid.). However, re-training the model with new data is costly in terms of time and computational resources, which is why OTT's core model has only been updated on a quarterly basis so far.

A sudden labor market shift occurred recently as a consequence of the COVID-19 crisis. The pandemic that hit 2020 brought severe disruptions in labor markets across the globe, skyrocketing unemployment rates and causing a shakeup in the job market (Dang and Viet Nguyen 2021: 2). Since then, studies have identified pandemic induced artifacts, for instance, that women were more likely to lose their job during the pandemic than men (ibid.: 6), (Kristal and Yaish 2020: 5). Besides women's employment being concentrated in the most affected sectors, another reason behind this effect was that, as schools and daycares were closed, many women suddenly had to take care of their children at the expense of working hours (Radulescu et al. 2021: 4). This relationship seems rather intuitive to a human interpreter. In hindsight, anyone with some understanding of labor market dynamics could have come to that conclusion. At the same time, it would have been completely impossible for OTT to preemptively capture this effect because it predicts exclusively based on past data-elicited relationships.

2.6 Formulating research objectives

Throughout this chapter, I have established that public sector ML-based decision support systems can and should theoretically be enhanced through integration with qualitative domain knowledge. I have presented a review of studies from various research fields where such endeavors have succeeded before and singled out a particular method that fits the general assumptions and use cases of decision support tools used in social and unemployment policy. The practical goal of this research is to adopt Gennatas et al. (2020) Expert Augmented Machine Learning method to discover to what extent can the predictive accuracy of a Machine Learning risk model be improved by modifying it according to experts' judgment. The broader aim of this research is to shed light on whether augmentation with expert knowledge is feasible and beneficial for ML-based tools in public services in general. For the purpose of guiding the analysis, a single concise research question is worded as follows:

What are the benefits of augmenting public sector data-driven systems with domain expert knowledge?

The expected benefits are threefold. First, an expert-augmented model is expected to solve or at least mitigate accountability- and ethics-related conundrums with public sector automation, as explained in previous chapters. It has been established that data-driven tools in this domain must contribute to fully explainable, reasoned, and ethical administrative decisions. It is of utmost importance that the internal formulae of Machine Learning models employed in the public domain complement the policies they are designed for. By relying fully on data-elicited relationships, an inductive statistical model can potentially make crucial mistakes by ignoring certain domain rules that the data does not adhere to, or historical biases that have been encoded into data and therefore will only be fortified by the model. Integrating expert judgment into a statistical model is a way of safeguarding against these pitfalls. Expert augmented Machine Learning models are theoretically able to capture and discard some of these hidden confounding artifacts that go into conflict with human judgment and domain rules.

H1: Expert-augmented models are able to reveal model artifacts that are in conflict with domain knowledge.

The second and most straightforward expected benefit is an increase in peak predictive accuracy. By using qualitative expert assessment to identify conflicting model artifacts, I effectively minimize model prediction error by eliminating features responsible for the least accurate risk estimations. As brought out in the previous chapter, a number of prior studies have shown that such approaches can yield significant accuracy benefits, especially when tested on data from different contexts. To measure the extent of this potential benefit, all models will be tested on three data sets with varying degrees of deviation from training data. The first data set will be drawn from the same distribution as training data to match its structure and underlying relationships. The second test set is comprised of individuals that became unemployed shortly after a sudden labor market shock, resulting in a visibly "abnormal" data structure. The third set includes data from a longer-term future time period, during which the labor market was expected to be recovering from said shock. The rationale behind choosing appropriate time frames for test sets as well as validation methods are explained in chapter 3.2.

H2.1: Expert-augmented models are able to outperform the base model in terms of predictive accuracy on previously unseen data with different structure and feature distributions compared to data used to train the model.

H2.2: Expert-augmented models yield the highest accuracy improvements on test data that differs the most from training data and the lowest on data that differs the least.

The third benefit that is expected as a direct result of this augmentation process has to do with model generalizability. As mentioned in chapter 2.4, a number of prior studies have discovered models that include some form of qualitative domain knowledge tend to perform better on less training data compared to their purely inductive counterparts[4]. A model that achieves good prediction accuracy on minimal training data would be useful in situations where 1) resources such as computational power and time are scarce – for example, when the model needs to be re-trained with new data at short intervals; and 2) labelled training data is costly or otherwise hard to come by. At first, it might seem like these aspects are not exactly at the forefront of priorities for OTT and other job-seeker profilers alike. There is usually no shortage of good input data, nor is it especially important to minimize the time and resources spent to re-train models. However, these benefits start to look more appealing when considering that frequent

[4]See for example Gennatas et al. (2020), Cao and Tay (2003), Schapire et al. (2002).

re-training can essentially contribute to the same goal as expert augmentation. Re-training ensures that the model is constantly adjusted to underlying shifts in the data structure that can possibly occur as a result of changes to domain rules or the surrounding environment. On the one hand, an important objective of expert augmentation is precisely to reduce the need to re-train models too frequently, as expert knowledge theoretically makes for a more versatile model that is better suited for different contexts. On the other hand, a model that is incidentally both expert-augmented and frequently updated with the most recent data can combine the best of both worlds. Constant re-training ensures that the model accounts for temporal changes, while integrated expert knowledge safeguards that the model is not overfitted to specific cases that are not in line with existing domain rules and common sense.

H3: Expert-augmented models will be able to achieve better prediction accuracy with less training data compared to the base model.

3 METHODOLOGY

The thesis at hand focuses on applying an existing knowledge engineering technique to a case from a particular field (social and unemployment policy) while adjusting it to meet the demands and assumptions of that field. While the former part of this problem setup strongly hints toward **applied research**, the latter ensures that this thesis also makes an academic contribution to the novel subfield of Machine Learning in the public sector. Indeed, research concerning Machine Learning is usually neither purely theoretical nor applied but instead falls somewhere on a smooth continuum between the two research types (Provost and Kohavi 1998: 128). A major cornerstone of the presented methodology is the Expert Augmented Machine Learning algorithm designed by Gennatas et al. (2020). Their multi-step approach, originally designed and tested in medical studies, will be optimized and validated for social and unemployment policy use cases. The ultimate goal is to shed light on whether this method is suitable for integrating human knowledge with public sector ML-based support tools in general.

The bulk of analysis, including the development and augmentation of models, was done in the R (v4.1.1) programming language using RStudio (v1.4.1717), a widely popular data analysis freeware. The R script used to produce the results of this thesis can be examined on the author's GitHub page[5]. All models were trained using their respective functions from the *rtemis* package (v0.83) developed by the authors of Gennatas et al. (2020) study.

3.1 Augmenting a public sector Machine Learning tool with expert knowledge

Gennatas et al. (2020) method was chosen as it allows to complete three crucial stages of domain knowledge integration while maintaining a good balance between research validity and feasibility. These four pivotal stages are listed below, together with brief explanations of the methodological steps taken to execute them in practice. Each step will be warranted in detail further along in this chapter.

1. **Transform a complex ML algorithm to a more easily interpretable rule-based format.** Train a replica of the OTT Random Forest model with individual level job-seeker data obtained from the Estonian Unemployment Insurance Fund. In parallel, train another model on the same data using a Gradient Boosting based RuleFit model as proposed by

[5] *https://github.com/peeterleets/expert-augmentation*

Friedman and Popescu (2008). This method allows to extract human-readable and easily interpretable decision rules from complex decision trees and tree ensembles. Specifics of this step are elaborated in **subchapter 3.1.1**.

2. **Design a platform for human experts to assess and validate the output of the rule-based model. Juxtapose empirical model-calculated risks with expert risk assessments and combine them for a theoretically better expert-augmented ML model.** In the same vein as Gennatas et al. (2020), have labor market experts qualitatively assess the risk of job-seeker subpopulations defined by each decision rule using a specially designed survey method. Calculate the difference between experts' assessments and empirical risks extracted from the RuleFit model. A specialized penalty value will be calculated for each rule based on 1) the difference between empirical and expert-assessed risk and 2) the extent of inter-expert variation for the expert assessment. Decision rules with the highest penalty values (biggest conflict between empirical estimation and expert judgment) will be successively removed from the final expert-augmented models in the hope for a better, more realistic model. This step is explained in detail in **subchapter 3.1.2**.

3. **Test the predictive performance of the expert-augmented model to confirm the benefits of integrating domain knowledge with Machine Learning.** Finally, the performance of the expert augmented model will be compared to that of 1) the Random Forest base model and 2) the base RuleFit model without expert calibration. All three types of models will also be tested with new, previously unseen data with slightly different underlying feature distributions.

3.1.1 Base model specification

In order to measure the improvement this methodology is able to offer to OTT, a base model replica is required to benchmark against. The first step in this methodology is to train a Random Forest base model that as accurately as possible represents OTT as it is currently employed by *Töötukassa* for profiling job-seekers. An ideal model replica assumes the availability of the very same data that was used to train the real model and that is as comparable in structure and scale as possible. A classification Random Forest benchmark model is trained with appropriate data obtained from the Estonian UIF. Hyperparameters for training the base model are chosen to match those of the OTT model in use. The number of trees trained for the ensemble (*ntrees*)

is 500, and the number of random features that are considered for splitting at each tree node (*mtry*) is automatically tuned for best model performance within the range of 4 to 15. The binary dependent variable, as defined by OTT, is whether or not the observation exited unemployment status within 180 days from registering as unemployed. While this is, in principle, a classification problem, the direct output of the model is interpretable as **the probability (0-1) of not exiting unemployment status within the next 180 days after registering as unemployed**. It is important to emphasize that the output is indeed the the inverse probability of resuming work in 180 days, interpretable as the *risk* of the positive outcome not happening[6]. This output definition is consistent across all models trained throughout this analysis.

Since, in theory, prediction rule ensembles can incorporate both Boosting and Random Forest learners under the hood (Fokkema 2020: 3), the next step is to determine which model type can yield better prediction accuracy on this data set to begin with. Accordingly, a competing Gradient Boosting model is to be trained with hyperparameters *shrinkage* (0.01) and *interaction depth* (10) chosen to reflect a good balance of model accuracy and interpretability[7]. Boosting hyperparameters were manually optimized so that, in the event GBM was to be chosen over Random Forest, the resulting ensemble would produce a realistic number of decision rules with enough predictors involved. This combination of shrinkage and interaction depth ensures that the resulting set of decision rules is comprehensive enough to contain enough useful information yet compact in size, which is crucial in regard to the feasibility of this study. In addition, it has been found that models with a small non-zero shrinkage parameter and a limited depth yield the best results overall (ibid.: 5). The number of trees to be trained is again 500 as it has shown to yield the best results for this type of base learner (Friedman and Popescu 2008: 926).

After determining the more potent learner, a RuleFit model will be trained to extract the most important prediction rules from the base model. The principal difference between RuleFit and regular tree ensembles considers its base learners. For tree ensembles, the base learner is an individual decision tree (Figure 1) that can be confounding and time-consuming to analyse. For

[6]As of May 1, 2022, OTT has been reconfigured to output the probability of resuming work instead. In this thesis, the previous definition is used as all analysis was completed by the time of this revision.

[7]It is important to note that the naming of the *interaction.depth* parameter for the GBM algorithm used in this study is slightly misleading. It defines the total number of splits done when growing a tree, not the actual depth of the resulting binary tree. Ergo an *interaction.depth* value of 10 means that starting from the root node, data is split ten times, resulting in exactly 11 terminal (leaf) nodes.

RuleFit, however, each base learner is in the form of a conjunctive decision rule

$$r_m(X) = \prod_{j=1}^{n} I(x_j \in S_{jm}), \tag{1}$$

where $I(\cdot)$ indicates the truth of its argument (Friedman and Popescu 2008: 919). Therefore, given S_j is the set of all possible input values, for any value $x_j, x_j \in S_j$, each rule $r_m(X)$ can either return 1 (*TRUE*) or 0 (*FALSE*) depending on whether the value belongs to its respective subset S_{jm} (ibid.). For continuous and ordinal variables, these subsets $S_{jm} = (t_{jm}, u_{jm}]$ define lower and upper boundaries $t_{jm} < x_j \leq u_{jm}$, and for nominal variables the subsets are explicitly enumerated (ibid.). In plain language, every decision rule captures a subset of the whole population based on some empirically defined criteria. For example, a decision rule

$$\textit{Gender} = \textit{male} \;\&\; \textit{Age} \geq 35 \;\&\; \textit{Age} < 20$$

would return 1 (*TRUE*) for observations that are male and between 35 and 20 years of age, and 0 (*FALSE*) for everyone else. The generation of those rules is just as intuitive – for each resulting decision tree, each node apart from the root node (ibid.: 920) is transformed into a rule based on which splitting variables were used to reach that node, as illustrated in figure 2.

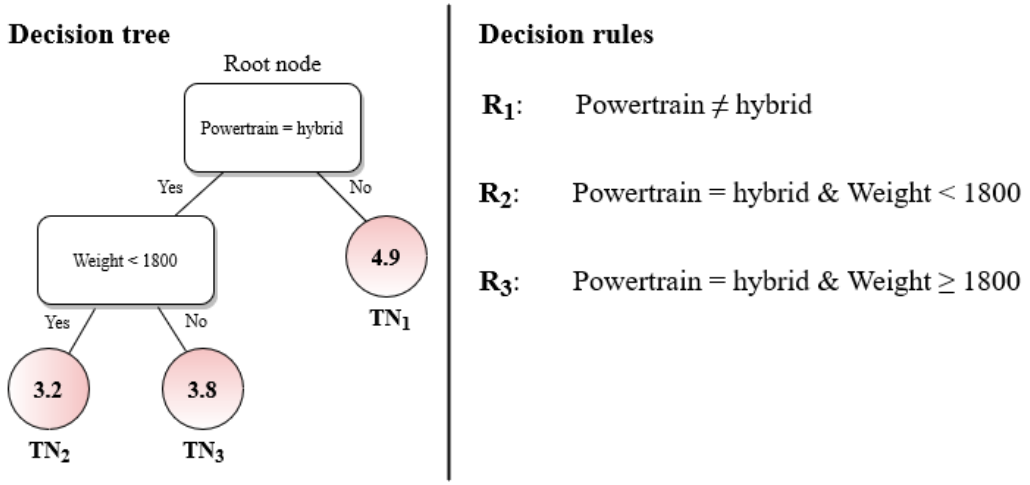

Figure 2. Example of a decision tree with three terminal nodes achieved by splitting data twice and its equivalent set of decision rules.

After the rule generation process, the resulting decision rules will be used in the final model as linear base learners with estimated variable coefficients (Fokkema 2020: 5–6). Transforming each tree node to a decision rule yields that the total number of rules K is

$$K = \sum_{m=1}^{M} 2(t_m - 1), \tag{2}$$

where t_m is the number of terminal nodes for the *m*th tree (Friedman and Popescu 2008: 921); and the total number of rules for terminal nodes is $M \times t_m$. Given a binary tree with 11 terminal nodes, a model trained with 500 such trees would yield a whopping 5500 decision rules. While some of these rules are likely duplicates (all 500 trees are grown independently from one another but are expected to split data in a relatively similar pattern), this number hardly supports the promise of a more interpretable method as it is practically infeasible to examine and comprehend the characteristics of over 5000 different subgroups. This is solved by extracting and including in the final linear model only these rules that significantly contribute to predictive accuracy (Fokkema 2020: 6). By default, RuleFit employs the LASSO penalty, which penalizes model terms with low predictive power and coefficients close to zero (ibid.; Friedman and Popescu 2008: 928). The LASSO shrinkage parameter λ that determines the severity of penalization applied to model coefficients is tuned for best training set performance via cross-validation.

While the author acknowledges that the ideal scenario would be to have experts assess each and every rule that obtains a non-zero coefficient, time and resource constraints can render that infeasible if the amount of selected rules grows too large. This problem is especially acute with the particular application under analysis. There are over 30 distinct predictors involved with the OTT replica trained in this study, with more than half of them being nominal. This means that reviewing each rule is expected to be more time consuming for experts than in Gennatas et al. (2020: 4573) study, which already reported an average response time of 41 ± 19 minutes per 126 questions with 3-5 variables each. To ensure the feasibility of this study with the resources at hand, it is needed to reduce the number of rules that experts will have to assess to a reasonable amount. In this case, ten LASSO-extracted rules with the highest and ten with the lowest model coefficients were selected for further expert validation. This choice of rules was somewhat arbitrary as, to my knowledge, no study has treated only a selection of model parameters with expert validation before. Given RuleFit parameter coefficients represent the extent of change in predicted value if the associated rule is satisfied (ibid.: 940), rules with the highest coefficients were considered to be the most influential in regard to predicted output values. This leads to the assumption that targeting these rules can potentially yield the biggest overall performance improvement.

3.1.2 Collecting and integrating qualitative expert risk assessments

The next step is to elicit expert assessment for rules selected by the LASSO-penalized RuleFit model. For this, a questionnaire was created using the University of Tartu LimeSurvey online survey tool. The questionnaire was comprised of as many questions as many rules were extracted from the RuleFit model, with each question corresponding to one rule. For each rule, experts were asked to assess the probability of a subpopulation of job-seekers to resume work in the next 180 days after registering as unemployed[8]. Experts had to assess the probability of a particular subpopulation defined by its respective rule relative to the entire population of job-seekers, also interpretable as the "average" job-seeker. Variables that defined a particular subpopulation (decision rule) were presented in the questions as a basis for assessment, similarly to Gennatas et al. (2020: 4573). A minimum of two and a maximum of five defining variables were presented for a subpopulation. For continuous variables, subpopulation arithmetic mean together with value range was displayed, whereas for ordinal variables, mode (most frequently occurring value) was provided instead. Equivalent indicators for the whole population were displayed for comparison. The response scale was a five-point Likert scale where the lowest response represented a much lower probability and the highest response a much higher probability for a person from some subpopulation to resume work, while the middle value represented probability equal to the whole population. Because some model-created decision rules can be too obscure for experts to make an informed assessment, a "cannot say" option was included as well. Moreover, experts were provided with a thorough yet concise guide for how to interpret and assess decision rules (see appendix 4). An example of a question is seen in figure 3. All 20 decision rules subject to expert assessment are listed in appendix 3, together with associated variables and statistics presented in the questionnaire.

The final and most important step is to integrate collected expert assessments with the final model. In the same vein as Gennatas et al. (2020: 4572), average expert assessment was calculated for each subpopulation. In their study, all LASSO-extracted rules were ranked from highest to lowest twice – one ranking was formed on the basis of model-calculated empirical risks and another on the basis of expert-assessed risks. The delta rank measurement was then calculated as the difference of both rankings ($\Delta R = Rank_{Em} - Rank_{Ex}$) that effectively

[8] Although, at the time of writing, OTT and equivalent models trained in this study output the probability of this event *not* happening, experts were presented with this phrasing instead as it was deemed more straightforward and easier to understand. Their responses were then inversed for the purpose of the analysis.

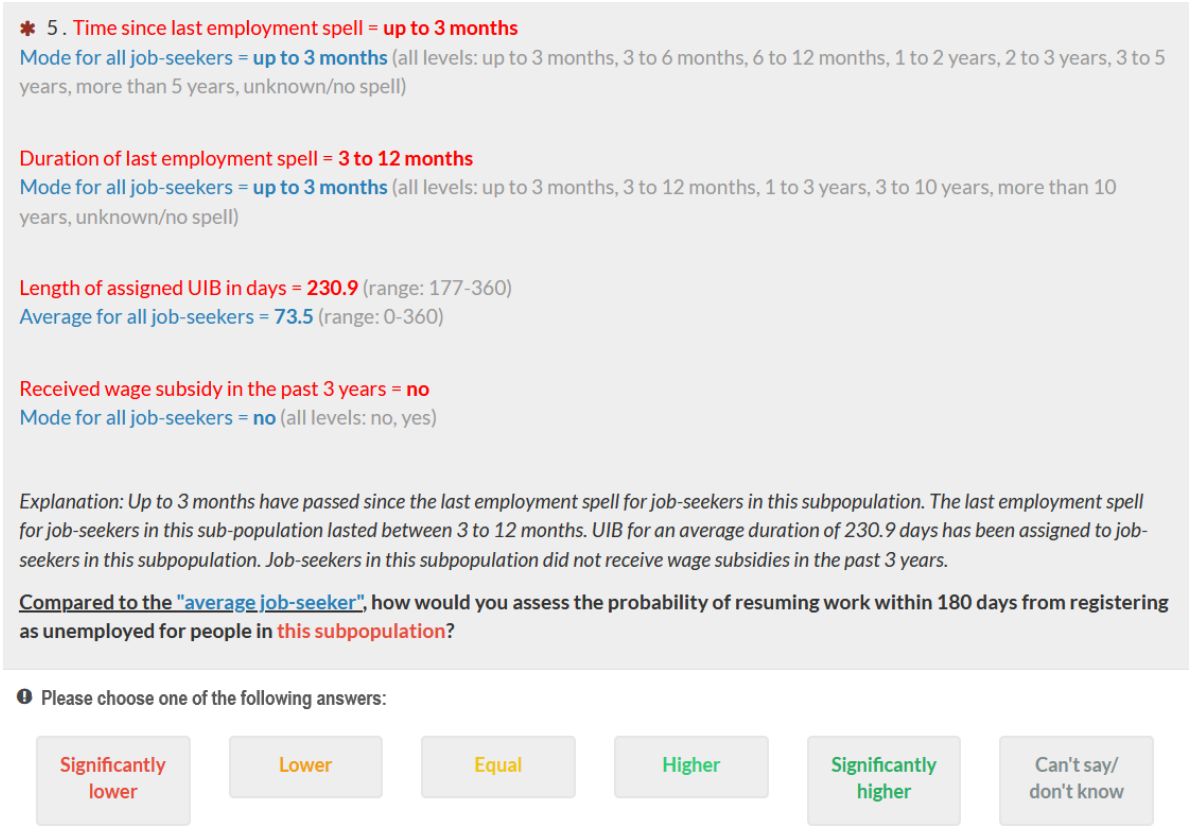

Figure 3. Example of a question presented to experts describing a particular subpopulation of job-seekers defined by a RuleFit-generated decision rule, translated from Estonian.

measures disagreement between experts and the empirical model (Gennatas et al. 2020: 4571). In the thesis at hand, however, a slightly different approach will be taken to detect and penalize supposedly unreliable rules. Instead of operating with rankings, the absolute difference between empirically estimated risks and expert-assessed risks will be calculated instead. Gennatas et al. (2020) method seems to be unsuitable for cases where multiple decision rules can have equal expert-assessed risks. In that case, the ranking of expert-assessed risks will be shorter than that of empirical risks because it will have two or more rules tied for the same rank. This means that rules further down the empirical risk scale end up having comparatively higher delta rank measurements and by extension, unfairly high penalties for the final model. Operating with absolute risks ensures that rules from both ends of the risk scale will be penalized on equal terms. To integrate inter-expert disagreement with the final penalty value P for rule r, the absolute difference between empirical and expert-assessed risks will be divided by the standard deviance for the average expert-assessed risk, such that

$$P_r = \frac{|Risk_{Em} - Risk_{Ex}|}{STDV_{Ex}}. \quad (3)$$

This yields that rules where experts disagreed with each other will be penalized less as the validity of expert judgment cannot be completely confirmed. Finally, rules will be binned into five ranks (R1-R5) based on their assigned penalty values to form five levels of penalization.

The final expert-augmented models take the form of a group-penalized regression where each linear term corresponds to a decision rule (thereby also including non-linear relationships modelled by RuleFit) that is penalized as a function of experts' disagreement with the empirical model and a measurement of trust in this disagreement in the form of $STDV_{Ex}$ (Gennatas et al. 2020:Appendix). Penalization is realized by excluding rules with penalty rank R higher than some predetermined threshold. In this case, a total of four expert-augmented models will be trained corresponding to each penalty rank minus the lowest (there is no reason to discard all expert-validated rules at once). The first model features the strictest penalization with complete discretion on expert assessment $R \leq 1$, and the last features all but the rules from the 5th R rank (only those rules where experts disagreed with the model the most).

Finally, prediction accuracy will be measured for all models using the AUC (Area Under the Curve) statistic, that "*represents the probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example*" (Huang and Ling 2005: 300). This measurement suits the use case of OTT well – since caseworkers are expected to use it primarily for ranking their clients from most to least likely to find a job, the numeric accuracy of the actual probability values themselves is not at the forefront, but rather is the correct ranking of those values. The AUC measurement has been used to evaluate classifier models in other related studies, notably Sinha and Zhao (2008).

3.2 Data sources and sampling method

This research utilizes two distinct types of data from both qualitative and quantitative research paradigms. The first data set, obtained from the Estonian UIF, was required to train base models that as accurately as possible replicate the actual in-use OTT prediction tool. Ideally, all 45 predictors used in OTT should be involved in the replica; however, due to strict data protection policies regarding sensitive personal data (for example, working capacity is a proxy to one's health condition) renders this practically impossible for research purposes. Therefore, significant aggregation was inevitable for certain variables. All such data transformations are described in

appendix 1. From all 45 predictors used in the real in-use OTT, 33 most important variables (according to model-calculated variable importance measure) were included in the replica to further reduce the possibility of identifying people in the sample.

The full data set consists of 423038 unemployment spells that were registered between January 1, 2015, and September 19, 2021 (both included). Four samples were drawn from this data set, feature distributions for each are presented in detail in appendix 2. For the purpose of training the models, a stratified random sample of 20% of all observations that were registered as unemployed **between January 1, 2015 and September 13, 2019** was drawn. The latter date refers to the latest point in time that one could have registered and still have 180 days to potentially exit unemployment before March 13, 2020, the day before the official COVID-19 lockdown period started in Estonia (more on that below). Three test samples from different time periods were drawn for model testing; the first test sample (hereinafter referred to as *test data*) was drawn from the same time period as the training set to match its structure and distribution. A stratified random 50% sample of all eligible observations was taken due to technical limitations for the size of the resulting feature matrix. Another test set (hereinafter referred to as *short-term future*) with theoretically different feature distributions stemming from the COVID-19 induced labor market shakeup was drawn from the period starting from March 13, 2020, to July 5, 2020 (all observations). A final test set (hereinafter referred to as *mid-term future*) represented a longer period of new unseen data, consisting of all individuals registered between March 13, 2020 and September 19, 2021 (both included). As seen in the table in appendix 2, COVID-19 induced labor market shifts manifest in almost every variable across data sets from different time periods. Some key shifts that could be observed are summarized below:

1. Since the beginning of the COVID-19 crisis, significantly more people who were recently employed suddenly became unemployed (see variables *time since last employment spell* and *status before unemployment*), often stemming from employer-related reasons such as lay-offs (see variable *reason for ending last employment*).

2. Proportionally more work-capable people with stable job histories suddenly became unemployed as indicated by *duration of last employment spell*, *months with payment in the last 2 years*, *work capacity*, *unemployment days in the last 3 years*, *previous unemployment spells*, *duration of assigned UIB* and *assigned UIB daily rate*[9].

[9]UIB with a higher daily rate for a longer period generally indicates that the individual has earned it with a long and stable job history (Töötukassa n.d.)

3. In regard to field of employment, the personal service sector took the most notable hit (as indicated by *field of last employment*), and it did not recover significantly even during the mid-term period.

4. For most variables, the greatest shift occurs between the base test set and the short-term future set, reflecting a sudden shock in the labor market caused by the COVID-19 crisis. In the mid-term future data set, full or partial recovery can be observed for many of these variables (see for example *months with payment in the last two years* and *number of people registered as unemployed around the same time*).

The domain knowledge to augment the models was sourced from a coalition of caseworkers from the Estonian UIF and a team of data scientists and labor economists responsible for designing and developing OTT. *Töötukassa* caseworkers were chosen for this study for two simple reasons. First, because most of them have already employed OTT in their day-to-day tasks, they need not be briefed about its goals and working principles in great detail, saving their as well as the researcher's time. Second, their particular knowledge and experience can also be regarded as most valid since OTT, a *Töötukassa* specific tool, operates with the exact data and features handled by them on a daily basis. Thirdly, as street-level unemployment officials, they have first-hand experience and knowledge of the Estonian labor market situation at any given time. The questionnaire described in subchapter 3.1.2 was distributed to 353 *Töötukassa* caseworkers, five of whom completed it in its full length. Two data scientists behind OTT were also included in the target group as their in-depth knowledge about the architecture and behaviour of OTT was deemed invaluable for detecting illogical model artifacts.

4 PRESENTATION AND INTERPRETATION OF ANALYSIS RESULTS

4.1 Training base models

A Random Forest OTT replica was trained to benchmark against RuleFit and expert-augmented models (AUC = 0.7218, table 3 in section 4.3). Random Forest offers stable prediction accuracy across the risk score distribution on test data with structure and feature distributions similar to training data, as seen on the top left plot in figure 4. However, the two plots to the right reveal obvious calibration issues at the extreme ends of the risk scale for new data drawn from future time periods. The middle plot represents prediction accuracy for individuals who registered as unemployed within the next few months after the start of the COVID-19 lockdown. It can be noted that Random Forest consistently overestimates the risk of staying unemployed beyond 180 days for individuals that in fact, did resume work in that time frame; and conversely underestimates risk for true high-risk individuals. This behaviour is also confirmed for the actual in-use OTT model. The rightmost plot for test data from a longer one and a half year time period indicates that, as the labor market situation recovers from the crisis over time, model predictions become more reliable again. Still, prediction accuracy has improved for low-risk individuals only, indicating that there is still room for improvement in terms of reliably detecting those people who will not resume work in the defined time frame.

A Gradient Boosting alternative was trained to determine which base model performs better on this data. As seen in table 3, Gradient Boosting consistently outperformed Random Forest on all three data sets. Despite overall accuracy improvements, the issue of miscalibrated risk scores that occurred with Random Forest persists with Gradient Boosting. As Gradient Boosting seemed to perform better overall, this model was chosen to facilitate the ensuing expert-augmentation process.

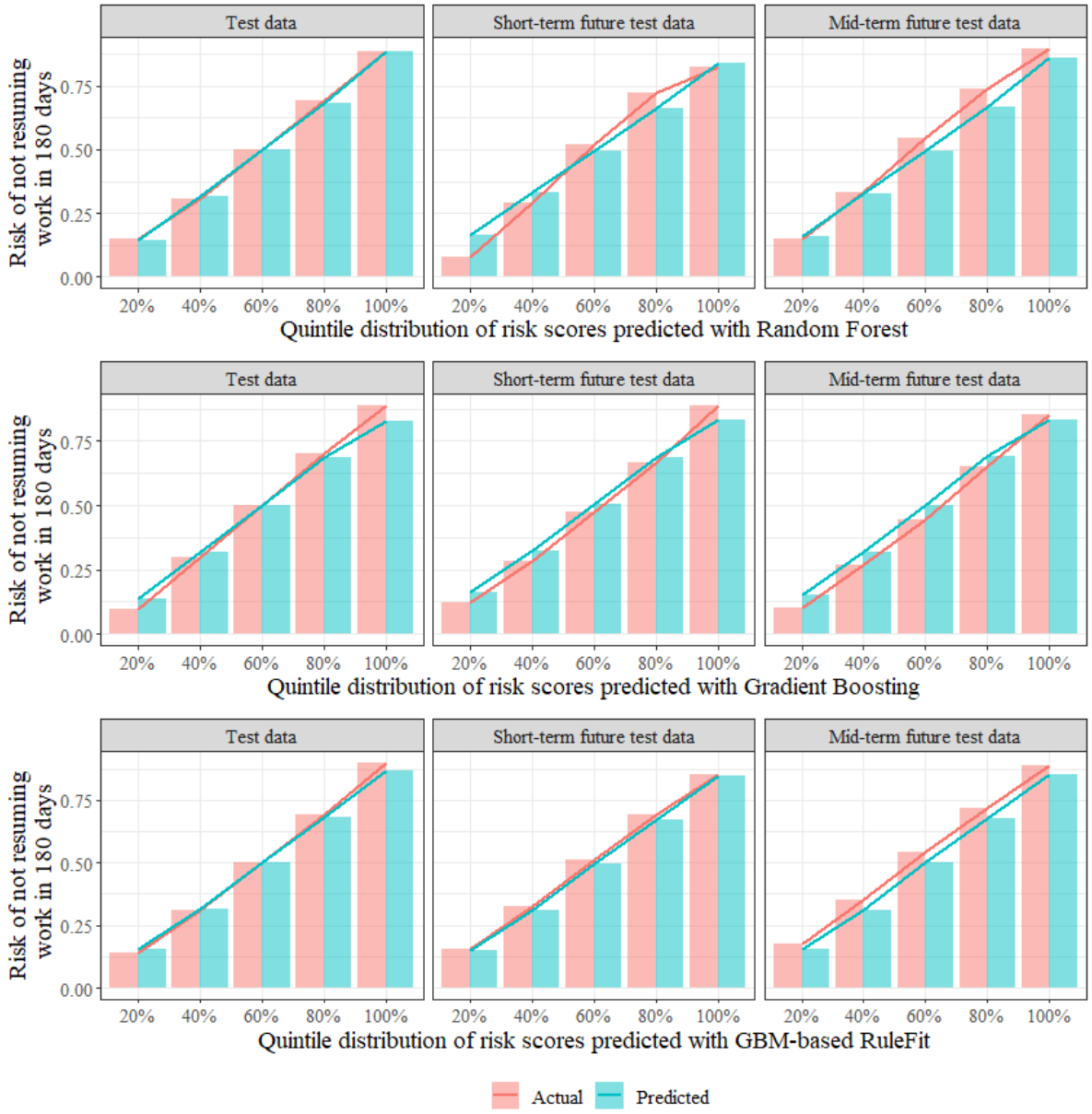

Figure 4. Quintile distribution of risk scores predicted with Random Forest, Gradient Boosting and GBM-based RuleFit. Observations were divided into quintiles based on their model-predicted risk scores (red bars). For every quintile, the average empirical risk – for how many observations the risk *actually* realized – was calculated for comparison (blue bars). Test data drawn from the same data distribution as training data vs test data from short-term and mid-term future time periods.

As seen in table 3 (chapter 4.3), the AUC accuracy measurements reflect the underlying data structures – all models yielded the highest AUC for the first test set that matched the distribution of training data; the lowest for data from a shorter time period immediately after the COVID-19 lockdown; and second-lowest for mid-term future data where labor market conditions were expected to have recovered to some extent. It can be noted that the LASSO RuleFit model with only the 142 most important rules was already able to outperform the RuleFit with all 3199

decision rules in terms of prediction accuracy on all test sets. Curiously, the GBM RuleFit model even outperformed the base Random Forest model on future data before it was augmented with expert knowledge, in spite of slightly lesser accuracy on the base test set (table 3). As seen in figure 4, RuleFit predictions for future data were also better calibrated, suggesting that this type of model is already more generalizable to new data and therefore may be less sensitive to sudden market shifts. From the 142 decision rules generated by the LASSO-penalized RuleFit model, 20 rules with the largest coefficients (absolute value) were selected for further expert validation. The resulting subset was mostly well representative of the initial rule set in terms of empirical risk distribution; however, on average, the final 20-rule subset included rules with slightly higher empirical risks assigned to them (figure 5).

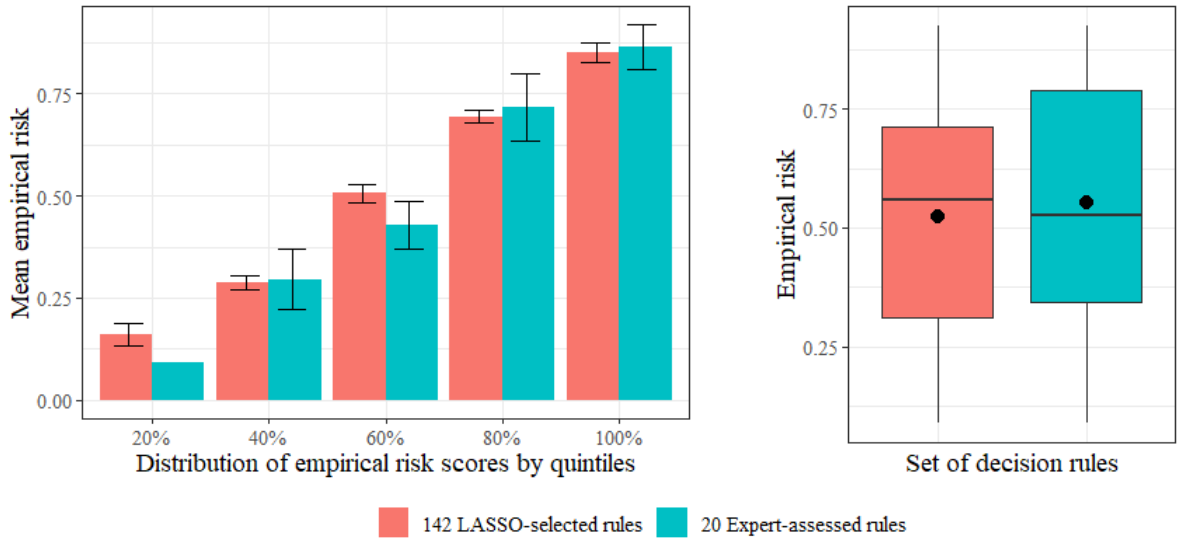

Figure 5. Distribution of empirical risk scores for the initial 142 LASSO-selected decision rules and the subset of 20 rules selected for expert validation. Error bars indicate 95% CI (left figure); black dot indicates sample mean and horizontal line sample median (right figure).

4.2 Validating collected expert risk assessments

A total of seven experts completed the questionnaire and assessed the likelihood of resuming work for subpopulations defined by selected decision rules. While this might seem like a small number of participants for a survey, relatively low standard deviation for responses indicates that, for most questions, experts' opinions did not clash very often, and elicited information would have likely started to saturate with a higher number of responses. Moreover, although Gennatas et al. (2020: 4571) application of this method involved 15 experts, other similar studies have

had to make due with insight from only as few as one single expert[10]. Expert-assessed as well as empirical risks for each of the 20 selected decision rules can be examined in appendix 3. The distribution of expert-assessed risks and empirically modelled risks can be seen in figure 7. It can be noted that the expert-assessed risks are distributed more evenly across the probability scale, which may be a good indication that 1) expert judgment differs from empirical induction to a certain extent, and 2) experts tend to give more pessimistic judgments to high-risk individuals and more optimistic judgments to low-risk individuals – confirming the potential to alleviate the miscalibration issue mentioned above and seen in figure 4.

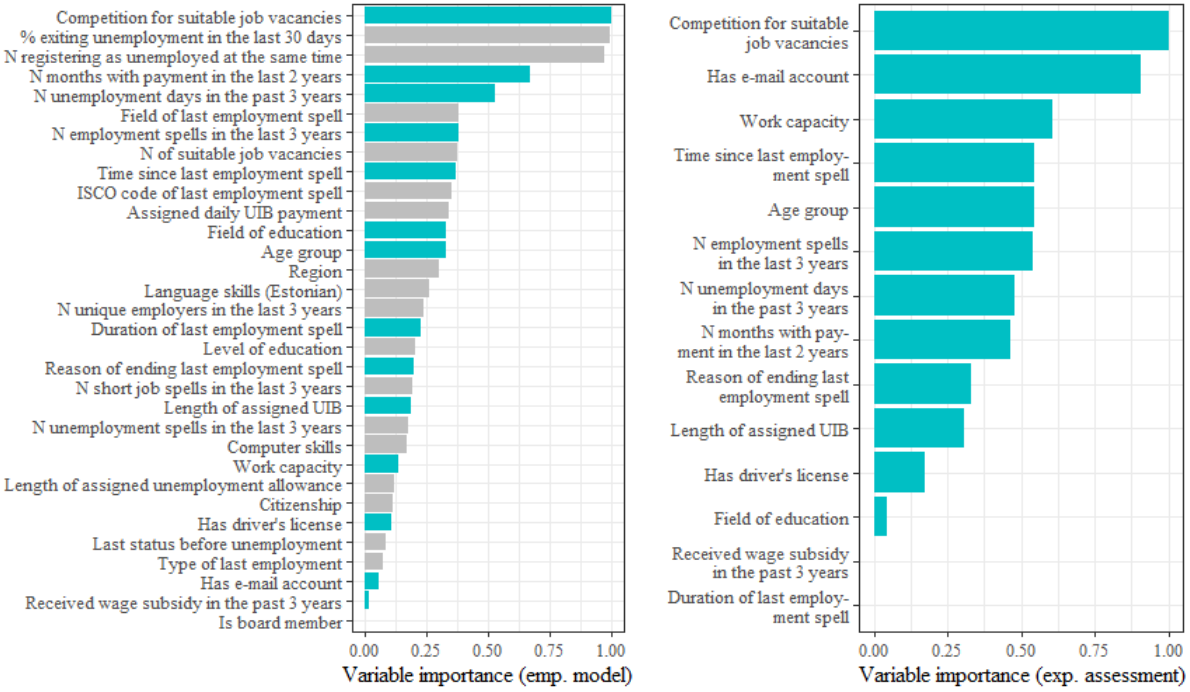

Figure 6. Model-calculated variable importance measure for individual predictors normalized to range 0-1 (left figure). Blue bars represent variables that also occurred in the 20 decision rules selected for expert validation. Constructed comparative variable importance measurement for expert assessments (right figure).

Figure 6 shows model-calculated variable importance of individual predictors for the Random Forest base model and a comparative estimation on the basis of experts' questionnaire responses[11]. First, it can be noted that among the most important predictors for the replica

[10]See: Sinha and Zhao (2008: 289); Schapire et al. (2002: 542)

[11]For every decision rule that includes a particular predictor, variable importance for that predictor increases by the absolute difference from the average response (the "equal" option equivalent to 0.5 on the probability scale) divided by the total number of predictors in the same rule. Each summed variable importance is then divided by how many times it occurred throughout the entire set of 20 rules to ensure correct scaling. Finally, calculated variable importance measures were normalized to a 1-0 scale.

base model were macro-level variables representing the overall labor market situation at the time training observations entered unemployment. This is where the base model trained in this study differs most from the real in-use OTT as for the real model, individual job history related variables tend to be most influential. This difference can likely be accounted to excessive aggregation of individual predictors and the exclusion of unique job-seekers in this study due to data protection requirements. Regardless of these inaccuracies and different approaches for calculating variable importance for empirical and expert assessments, the overall concurrence of these rankings helps to confirm the validity of collected expert risk assessments. Interestingly, whether an individual has a functional e-mail account seems to have been an important flag for experts. It can be hypothesized that, for experts, having (or providing) an e-mail account as a contact form proxies motivation to find a job as soon as possible. More likely, however, it ranks this high because it appeared in decision rules where other variables were already strong indicators of whether or not an observation is capable of finding a job fast. Compared to the base model, work capacity also ranks considerably higher as it does for the real in-use OTT model.

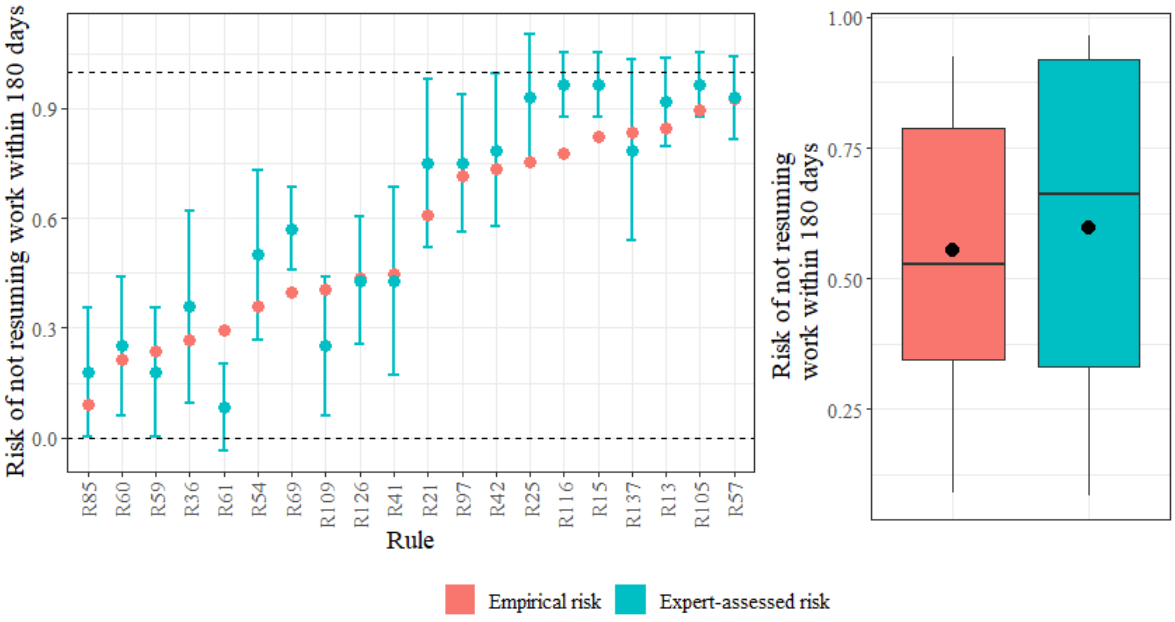

Figure 7. Empirical risks and expert-assessed risks for each selected decision rule (left figure). Error bars for expert-assessed risks indicate 95% CI. Distribution of empirical risks and expert-assessed risks for the set of 20 selected decision rules (right figure); black dot indicates sample mean and horizontal line sample median.

Expert-assessed risks and empirical risks were also plotted for each rule to ensure further that valid information was acquired from caseworkers (figure 7). A monotonic relationship exists between empirical and expert-assessed risks, confirming the general validity of collected expert knowledge. Despite this, five rules out of 20 fell outside the 95% confidence interval as seen in

figure 7, indicating a substantial difference between expert and model judgment. Following the methodology of Gennatas et al. (2020: 4573), these rules (presented in table 5) were further investigated as they may reveal hidden confounders or systematic expert misjudgment.

The first notable discovery is that out of these five, three (116, 15 and 25) included job-seekers from the same age group of 50-60-year-olds (mode for the "average" job-seeker being a much lower value of 20-30). All three such subpopulations fell on the higher end of the risk scale, and in every case, experts significantly overestimated the risk of staying unemployed compared to the empirical model. Additionally, out of 7 times the *age group* variable appeared throughout the rule set, *50-60* appeared exactly thrice, meaning that experts overestimated the risk for every single assessed rule defined by this age group. There is a good reason to suspect that this points towards a hidden confounding artifact. Contrary to what may sometimes be mistakenly stated, a probabilistic model cannot actually under- or overestimate the *empirical* risk of some event happening if it already knows the outcome. For the model to output a risk score of 0.776, this outcome indeed had to happen for exactly 776 individuals out of 1000 that satisfied this decision rule in the training data. The fact that qualitative expert judgment systematically differed from the inherently objective empirical risk suggests that experts either misjudged the risk or some unobserved confounding variables and/or *ex-post* interventions are secretly modifying it (Gennatas et al. 2020: 4573).

Assuming that expert assessment was exclusively based on logical deductions from factors present in decision rules, there are two possibilities as to why the risk scores are so different. The first possibility is that these particular subpopulations received some sort of intervention in the form of special unemployment benefits and/or training programs that the experts did not foresee. This intervention ultimately improved their chances of finding a job, effectively lowering their risk of staying unemployed. Another explanation could be that there are other confounding variables at play that result in the empirical risk being lower than it may seem from the hard variables present in that rule. I will bring a purely hypothetical yet plausible example of how a hidden confounder may be distorting the empirical risk score in this case. Besides all three subpopulations mostly consisting of 50-60-year-old job-seekers, it can be noted that they also hold individuals that *to our knowledge* have not worked for a relatively long time[12]. It can be hypothesized that these rules have incidentally also captured completely work-capable people

[12]See variables *time since last employment spell* and *n of employment spells in the last 3 years* for rules 116, 15 and 25.

who had recently returned from working abroad. Because there is no accurate track record of their job or travel history, experts may have mistakenly assumed that these people are less capable of finding a job than they apparently are; thus, the more pessimistic evaluation. It goes without saying that any such explanation is mostly speculative as they cannot be confirmed without additional relevant information. Still, this hypothetical confounder would explain why the average empirical risk for all three of those subpopulations was notably lower for short-term data – as travel restrictions imposed by the COVID-19 crisis were looming on the horizon, it seems plausible that more job-seekers had returned from working abroad than before. Speculations aside, the most important thing is that in this case, expert assessment theoretically offers a more *pure* risk prediction because it is not muddled by unobserved (and often unexplainable) confounding variables and interventions. Irrespective of the true explanation, there seems to be enough evidence that individuals between the ages of 50 and 60 that have officially been inactive for a while, can be considered a confounding artifact. Thus, hypothesis 1 can be confirmed.

Table 2. Five rules for which expert assessment significantly differed from empirical risk (expert-assessed risk fell outside the 95% confidence interval) ordered by assigned penalty value ($\frac{|\Delta Risk|}{STDV_{Ex}}$) from highest to lowest.

ID	Subpopulation defined by decision rule	Empirical risk	Expert assessed risk	Expert STDV	Penalty	R
116	Time since last employment spell = unknown/no spell Competition for suitable job vacancies = 0.03 Age group = 51-60	0.776	0.964	0.094	1.994	5
61	N of employment spells in the last 3 years = 1 N of months with payment in the last 2 years = 24 Time since last employment spell = up to 3 months N of unemployment days in the past 3 years = 1.1 Length of assigned UIB in days = 0	0.294	0.083	0.129	1.632	5
15	N of employment spells in the last 3 years = 0 Time since last employment spell = 3 to 5 years Age group = 51-60	0.824	0.964	0.094	1.483	4
69	Time since last employment spell = up to 3 months Duration of last employment spell = 3 to 12 months Length of assigned UIB in days = 230.9 Received wage subsidy in the past 3 years = no	0.397	0.571	0.122	1.427	4
25	Time since last employment spell = 3 to 5 years Age group = 51-60	0.754	0.929	0.189	0.924	3

4.3 Penalizing unreliable rules to form final expert-augmented models

Penalties were applied to decision rules based on the extent expert assessments coincided with the empirical risks assigned to each of these rules. These penalties ultimately determined which rules were to be included in the final expert augmented linear regression models. Figure 8 shows applied penalties for each expert-assessed decision rule. It can be seen that higher penalization occurs when 1) the absolute difference between empirical and expert-assessed risks is higher, and 2) the standard deviance for the recorded expert assessment is low (indicating inter-expert agreement). The 20 penalized rules are then discretized into five ranks R1-R5. Four expert-augmented models were trained with each subsequent model excluding another rank of decision rules from the set of predictor variables. For example, the first expert augmented model includes rules for which $R \leq 4$ (rules 61 and 116 will be excluded as seen in figure 8), the second model rules for which $R \leq 3$ and so on. Another model was trained by specifically discarding rules 116, 15 and 25 that were suspected to involve a common confounding variable, as discovered before.

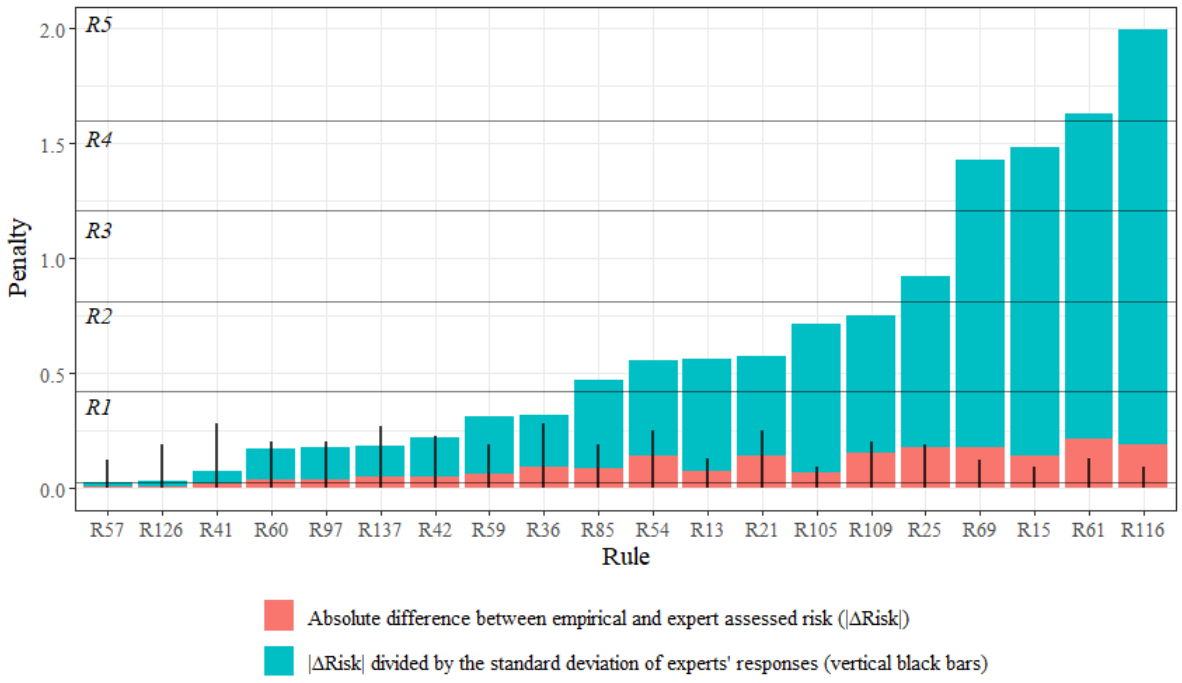

Figure 8. Penalty values for each decision rule subject to expert assessment. Red bars display the initial absolute difference between empirical and expert-assessed risk ($|\Delta Risk|$). Blue bars display the final calculated penalty value, obtained by dividing $|\Delta Risk|$ by the standard deviance for expert assessment (inter-expert agreement, represented by black vertical bars).

AUC accuracy measures for all four expert-augmented models are shown in table 3. On the base test set, none of the expert-augmented models was able to improve in accuracy over either the LASSO RuleFit nor the Random Forest model. On future data sets, expert-augmented models were able to consistently outperform the Random Forest benchmark model, however, no improvements were recorded over the base non-augmented RuleFit and boosting models. The one expert-augmented model that matched the accuracy of RuleFit on all but one test set (trailing by just 0.01 percentage points on mid-term future data) was the one that was specifically configured to exclude the rules with the aforementioned hidden confounder. Other relatively potent performers included model $R \leq 1$ that strictly only retained the most reliable rules and rule $R \leq 4$ that only discarded the most conflicting rules. Since no overall improvement in AUC was recorded for any of the expert-augmented models compared to a non-augmented RuleFit, hypotheses 2.1 and 2.2 cannot be confirmed.

Table 3. Comparison of the predictive accuracy of trained models based on test samples from 1) the same distribution as training data (job-seekers who registered as unemployed between January 2015 and September 2019); 2) a short-term future time period (March-July 2020); 3) a long-term future time period (March 2020 to September 2021).

Step	**Model**	**AUC (test data)**	**AUC (short-term future data)**	**AUC (mid-term future data)**
OTT replica model	Random Forest	0.7218	0.6836	0.6917
GBM base model	Gradient Boosting	0.7238	0.6975	0.7030
RuleFit	GBM-trained RuleFit (all rules)	0.7128	0.6879	0.6967
	GBM-trained RuleFit (142 LASSO-selected rules)	0.7131	0.6880	0.6967
Final expert-augmented models	$R \leq 4$	0.7129	0.6877	0.6965
	$R \leq 3$	0.7127	0.6876	0.6961
	$R \leq 2$	0.7126	0.6875	0.6961
	$R \leq 1$	0.7127	0.6878	0.6962
	Discarded R116, R25 and R15	0.7131	0.6880	0.6966
Average improvement of expert-augmented models	over Random Forest	-0.0089	**+0.0041**	**+0.0046**
	over Gradient Boosting	-0.0109	-0.0098	-0.0067
	over LASSO RuleFit	-0.0002	-0.0003	-0.0004

To investigate how this penalization method affected the model, predicted risks were plotted for ten (out of 142) rules for which the best expert-augmented model prediction differed most from regular RuleFit prediction, regardless of whether these rules were actually subject to expert assessment (figure 9). It can be noted that all such rules that saw their associated risk predictions improve (although incrementally) included individuals with either very high or very low empirical risks. That can be considered a sign that the resulting expert-augmented model has somewhat better generalization properties than the base RuleFit, predicting slightly more reliably for irregular data points. Moreover, five out of ten rules present in figure 9 formed the five smallest subpopulations in the model. This is further evidence of improved generalizability and reduced overfitting, as the resulting expert-augmented model has the strongest corrective effect on the most uncommon cases. It can also be noted that only one rule (R61) that itself was subject to expert validation was significantly affected by the expert-augmentation process. This suggests that predictive performance across the entire model can be affected by only having experts validate a fraction of its parameters.

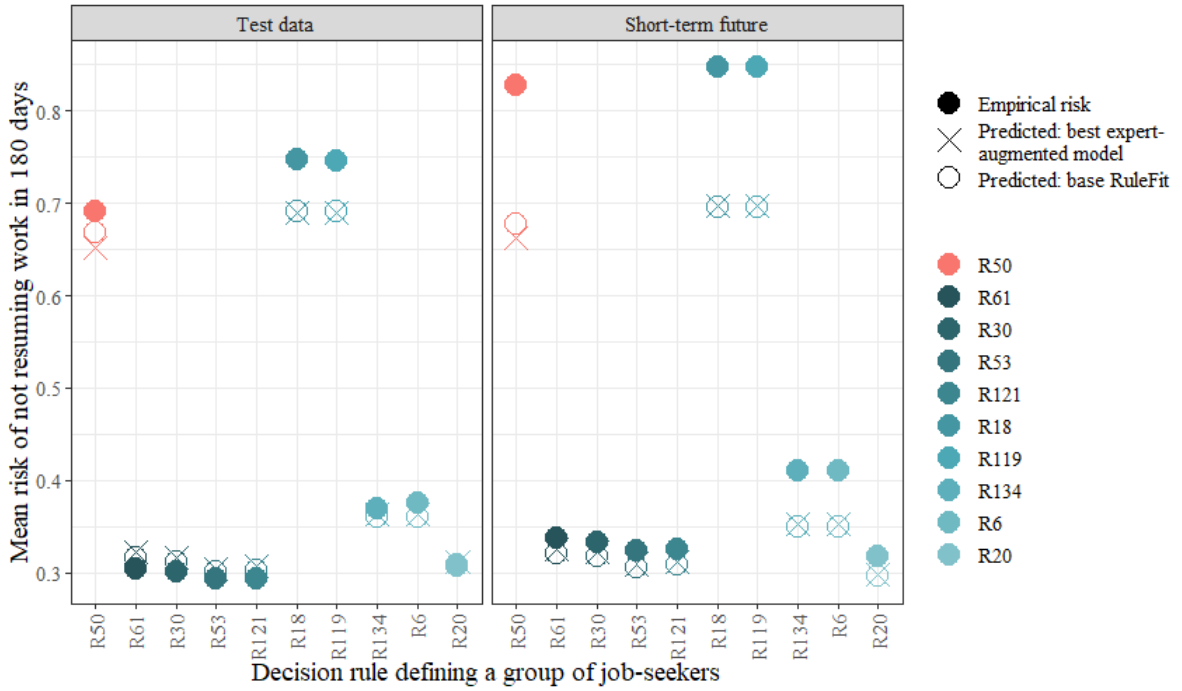

Figure 9. Average empirical and predicted risks for decision rules where the absolute difference between predictions given by the $R \leq 1$ expert-augmented model and the base RuleFit was the highest. Rules are ordered from highest to lowest in terms of absolute difference in predicted risk scores. Red represents rules for which the expert-augmented model predicted a higher, and blue rules for which it predicted a lower risk than the base RuleFit model.

While these improvements were indeed marginal, predictive accuracy notably declined for one decision rule. As seen in figure 9 in red, the best expert-augmented model predicted a significantly lower score for R50. Upon closer inspection, this rule was confirmed to capture similar individuals to removed rules R116, R15, and R25. It can therefore be assumed that, had this particular rule been a part of the expert-validated rule set, it would have also emerged as a confounding variable that should not be included in the model formula. The performance dip in the $\sim$70% risk range highlighted in figure 10 confirms that rule 50 largely accounts for the lesser overall predictive performance of the best expert-augmented model compared to base RuleFit. The same figure also suggests that rule-based models perform better at the higher end of the risk scale, although this artifact is likely unique to this particular data and/or subject of analysis.

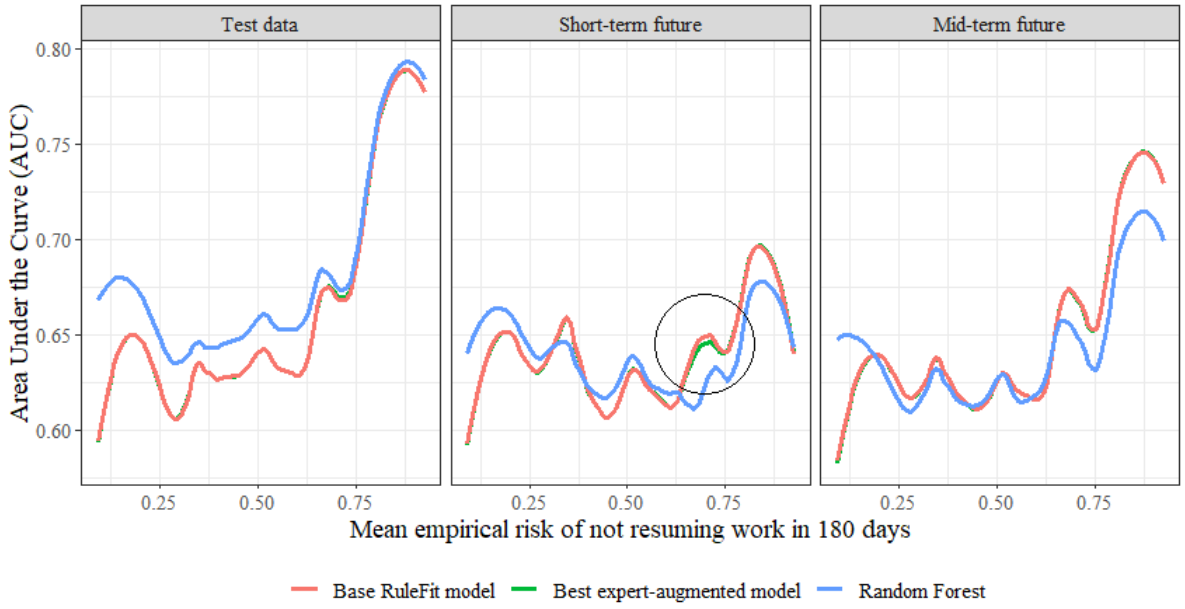

Figure 10. AUC measurement vs empirical risk for each rule-defined subpopulation – base RuleFit vs Random Forest vs the best expert-augmented model (discarded rules 116, 15 and 25).

Performance of all five expert-augmented models and the RuleFit base model was also evaluated on different training sample sizes. Figure 11 shows classification accuracy for all models trained on samples ranging from 500 to 16000 observations in size. It can be seen that with less training data, two expert-augmented models, $R \leq 4$ and the one that discarded a suspected confounder, performed marginally yet statistically significantly better than base RuleFit. As expected, this improvement was the most notable on short-term data – the confounder-discarding model learned faster up to about 4000 training observations. Hypothesis 3 can therefore be partially confirmed – while expert-augmented models proved to learn slightly faster for short-term data and base test data, no improvement was recorded for mid-term future data. Furthermore, the

improvements were relatively marginal; however, that is likely due to the fact that only 20 rules out of 142 received expert validation. My results suggest that there is potential for considerably faster-learning models with more extensive integration of expert knowledge.

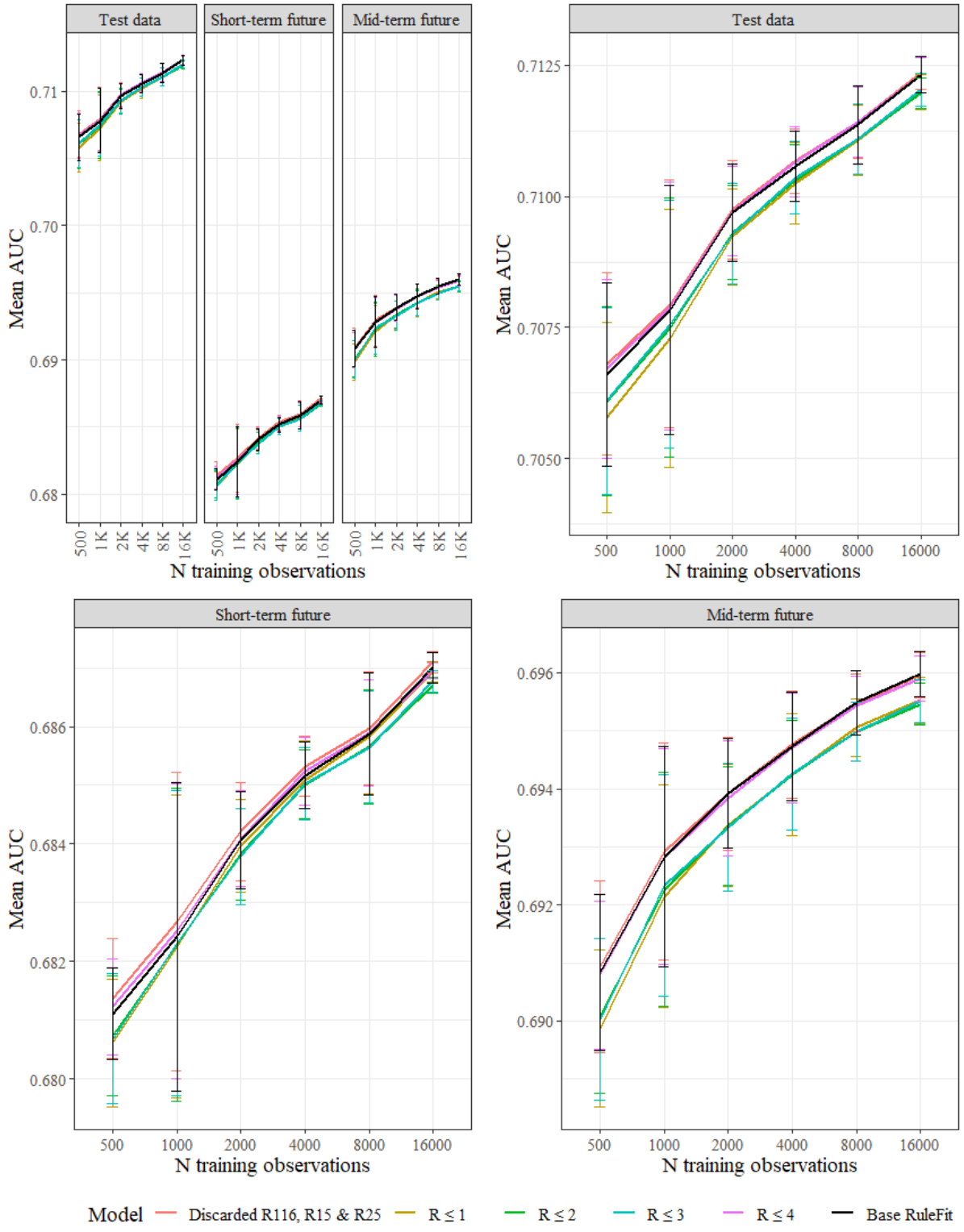

Figure 11. Predictive performance of expert-augmented models vs base RuleFit on different size training sets. Error bars indicate 95% CI across five stratified random subsamples. Top-left plot represents a "zoomed out" view.

4.4 Discussion

As more and more tasks are outsourced to automated tools in this sector, the level of accuracy these machines can perform public tasks with is becoming under scrutiny. From a policy standpoint, the accuracy of statistical tools contributes to the overall legitimacy of the administrative procedures they are integrated into. Given the legitimacy of state agencies hinges on the competence of their internal workings, it is crucial that these processes are fully accountable to the public. Introducing complex machines in this accountability chain certainly makes it more difficult to trace the roots of administrative decisions and actions. This puts further weight on prediction accuracy and reliability as the output of these machines must be in compliance with policy goals and the surrounding context at all times.

Augmenting public sector Machine Learning systems with expert knowledge was expected to yield three benefits in this area. First, integrating human experience and value judgment with Machine Learning models was to boost its predictive performance on new data (*H2.1* & *H2.2*). Although this analysis was unable to confirm substantial accuracy improvements in augmenting Machine Learning models with expert knowledge, certain findings give reason to believe that the expected benefits discussed in literature hide right around the corner. It was expected that expert-augmented models are more generalizable for new contexts, resulting in better predictive performance on fewer training data (*H3*). Performance evaluation of models showed that some expert-augmented models can learn slightly faster with less training data compared to purely quantitative models, indicating better generalization properties on out-of-sample test data with alternate structure and feature distributions. While the recorded improvements were relatively marginal, they nonetheless harmonize with other similar studies, most notably the work of Gennatas et al. (2020) whose approach was adopted in this thesis.

Somewhat unrelated to the primary goal of this research – to improve the accuracy and generalizability of ML-based tools through integration with expert knowledge – another potential benefit emerged during the base model preparation phase. The particular expert-augmentation method followed in this study required transforming a complex tree-based ML model to a more simplistic and intuitive rule-based format. Despite the fact that this step did not yet involve additional qualitative data, the resulting prediction rule ensemble model was able to outperform the tree-based benchmark model as well as each subsequent expert-augmented model in terms of prediction accuracy. As some previous studies have demonstrated that prediction rule ensembles can yield superior classification accuracy compared to more complex tree-based

models (Friedman and Popescu 2008: 926–927), this finding was not completely surprising on its own. What is noteworthy, however, is that these types of models are decidedly easier to comprehend due to being composed of a relatively small number of independent linear components interpretable as decision rules (Fokkema 2020: 1–2). Contrasting different models from the perspectives of interpretability and explainability was ultimately not in the scope of this study. Regardless, a reasoned assessment can be made based on the extracted set of decision rules that this type of model allows for a more comprehensible examination of different variables and latent factors that influence one's risk score. This is supported by some caseworkers' positive feedback to the rule assessment questionnaire (collected independently from risk assessments for validation purposes), highlighting the straightforwardness of presented decision rules. Using prediction rule ensembles instead of tree-based models might therefore help ML tools meet the explainability requirement that is crucial in the public domain. Despite having not been a direct objective of this study, this *posteriori* knowledge opens up interesting avenues for future research in the field of explainable AI and Machine Learning.

Increased predictive accuracy was not the only improvement achieved by transforming a tree ensemble model to a simpler, rule-based model. Comparison between trained Random Forest and RuleFit showed that the latter was able to offer more stable prediction accuracy throughout the risk scale on new data. Random Forest proved to miss the mark, especially at the far ends of the scale, often yielding overly optimistic scores for those at risk and pessimistic for those not at risk. A rule-based model proved to remedy that by having prediction errors distributed across the scale more evenly. For policy areas where the actual risk scores are of little relevance and the focus is on how these scores ultimately rank individuals (such as job-seeker profiling), the calibration issue innate to Random Forest is not a deal-breaker. However, as indicated by this finding, prediction rule ensembles might be a good fit for use cases where numeric differences in predicted values are also meaningful from a policy perspective.

While this thesis empirically tested improvements to a public sector Machine Learning model from pure predictive performance and domain logic consistency standpoints, the literature dissected in chapter 1 suggested another theoretical benefit to this endeavor. Data-driven systems tend to become more interpretable and explainable when infused with some form of prior knowledge. The premise for this was that as expert knowledge helps to discover confounding artifacts in the form of unobserved factors and interventions, models that discard those artifacts become more easily explainable since their output is more consistent with real-life scenarios

(*H1*). In this analysis, systematic differences in expert and model assessments revealed that for job-seeker groups with certain common observed features, a confounding latent variable was modifying model-predicted risks. Model parameters involving this unexplained confounding variable were then penalized in the hope of a more accurate model. While, in this case, some statistical evidence emerged that it may have indeed been a hidden confounder (as suggested by increased generalization properties of the respective expert-augmented model, including the ability to learn faster on new data), the true explanation behind it can only be speculated. It was not in the scope of this thesis to collect additional expert opinions regarding discovered confounding variables to validate whether they are consistent with domain context. Future studies that apply this method can improve the validity of results by conducting another round of expert deliberation in the form of a survey or a focus group interview to confirm that the exclusion of this unobserved factor is indeed necessary and justified.

There are other limitations to the results of this study that need to be addressed. The most prominent bottleneck of this analysis was that only a fraction of all model components were given the expert knowledge treatment. Due to resource constraints, only 20 decision rules out of 142 were chosen for expert validation, meaning that the hypothesized accuracy improvement should have manifested itself in less than 30% of all model components. I identified and discarded what was deemed to be a hidden confounder according to differences in model and expert risk assessments. The problematic part is that these differences are, in part, relative to the set of rules chosen for expert validation. It was my analytical decision to have experts assess exactly 20 rules with the highest model coefficients to maximize expected improvements. A differently defined set would have possibly revealed different confounding variables. It can also be that experts judged each rule not entirely based on provided population statistics but rather on how they compared to other rules in the questionnaire, resulting in somewhat distorted assessments specific to the particular set of rules subject to qualitative validation.

Another limitation concerns a trade-off between research feasibility and the quality of elicited information. In this study, the questionnaire for collecting expert risk assessments was deliberately designed to be as easy to comprehend and respond to as possible. Although prediction rule ensembles are often lauded for their interpretability benefits compared to more complex algorithms, some model-created rules can still end up being hard to comprehend or outright nonsensical. While the author could have gone the extra mile in explaining how to interpret each rule, a conscious decision was taken to avoid cluttering the questionnaire with

long instructions and excessively detailed questions. The reason for that is simple – answering 20 questions that each require special attention to subpopulation statistics and variables is time-consuming and, frankly, rather boring. As the author's resources to compensate for respondents' efforts were limited, compromises had to be made in regard to how much information was presented to them. One such compromise was that, unlike in Gennatas et al. (2020) study, only population mode was displayed for nominal variables, with no reference to other levels that were also present in that subpopulation. For example, a group of job-seekers was presented to be construction workers by training, despite that actually being the most frequently occurring level in that subpopulation. Other levels were not displayed to avoid overcomplicating the questions and potentially confusing respondents. It would have been hard to make an educated risk assessment if that group was said to consist of individuals from many seemingly unrelated fields – for example, a mix of construction workers, lawyers, and family doctors. On the other hand, as confusing as some of these combinations may have been, presenting them exactly as they appeared in model-created decision rules may have contributed to revealing confounding variables. As a total of seven experts completed the questionnaire – just above the desired minimum – the decision to sacrifice some information quality for a satisfactory number of respondents ultimately proved to be the right call for this study. Future research using this approach can expect to elicit better quality expert knowledge with more detailed questionnaires that perfectly represent all model parameters.

Finally, the broader aim of this thesis was to generate knowledge regarding the technical benefits of expert-augmentation for public sector machine learning systems in general. While this thesis successfully demonstrated the viability of said approach within overarching implementation constraints that exist in the public domain, these concrete findings are still largely particular to decision support tools in social and unemployment policy. The set of decision rules to be validated by experts, as well as the AUC accuracy measurement for benchmarking were chosen with the specific objectives of OTT and welfare provision systems in mind. As OTT can ultimately only determine the order in which policy-based intervention is to be delivered among the society, the correct ranking of predicted values was ultimately in the crosshairs. It cannot be taken for granted that these findings stand for use cases with completely different priorities and definitions regarding model *goodness* – for example, cases where predicted nominal values yield greater policy implications instead.

5 CONCLUSION

The public sphere is a notoriously challenging domain to design automated systems for. Implementation barriers include miscommunication issues between technical developers and public procurers, high system performance and cost requirements stemming from the ramifications of administrative decisions, and ethical considerations in relying on data-driven systems to steer human lives. The aim of this applied thesis was to explore how public sector machine learning systems can be improved through integration with qualitative expert knowledge in order to mediate some of these concerns. First, qualitative domain knowledge has proven to enhance the predictive performance of machine learning models, especially on new data. Second, by augmenting data-driven models with qualitative expert input, system integrity and consistency with domain rules can essentially be validated, resulting in a model that fits its intended use case and policy objectives better.

The subject for this study was OTT – a real in-use ML-based decision support tool from the field of unemployment policy. The tool in question estimates long-term unemployment risks of job-seekers based on quantitative individual-level data, including a fairly large number of standard socio-demographic and job history variables known to affect one's labor market status. Training data relied on by the model inherently reflects the labor economy situation at the time of its collection, meaning that sudden labor market changes in the temporal dimension can render this model ineffective. A more flexible model with better generalization properties was expected to remedy that issue. Therefore, the aim of this study was to test whether expert-augmentation 1) improves the overall predictive accuracy of a public sector machine learning decision support tool, 2) yields a model more generalizable for new data, and 3) helps to identify model artifacts not in line with the rules and objectives of its use case.

A suitable knowledge engineering method was adapted from medical studies to incorporate domain knowledge with a data-driven model. This method entailed transforming a complex ML model to a collection of simple decision rules that were then subjected to expert validation. System developers and caseworkers from the Estonian Unemployment Insurance Fund qualitatively judged the risks of 20 rule-defined job-seeker groups through a specially designed platform. Model-calculated risks were then compared to expert-assessed risks, and those decision rules with substantial differences in empirical and expert risk estimations were removed from the final expert-augmented models.

This study was unable to confirm substantial peak accuracy benefits of the tested approach. Although the procedural step of converting a complex ML model to a more interpretable rule-based format did improve overall accuracy, the ensuing expert-augmentation process itself did not. However, expert-augmented were confirmed to be more flexible as they 1) performed comparatively well on new data, 2) could learn faster with less training data, and 3) yielded somewhat higher prediction accuracy for uncommon observations. Moreover, this method was able to reveal job-seekers for whom empirical and expert risk assessments systematically differed, indicating the presence of a common hidden confounding variable. Although this study did not validate the consistency of that confounder with domain logic, it was confirmed that its exclusion from the model contributes to increased generalization properties.

These findings yield important implications regarding future use cases for Machine Learning based decision support tools in the public sphere. As was brought up earlier, a myriad of implementation challenges, including strict data regulations, cooperation barriers between procurers and system developers, and budget constraints, hinder advancements in public sector automation. For one thing, models with good generalization properties can be deployed in various different contexts with no significant drop-off in predictive performance. My analysis provides some evidence that machine-learning models augmented with expert knowledge can adapt to irregular contexts better. On the other hand, systems that can learn with less input data are not only more versatile but can yield serious cost optimization benefits in areas where labelled training data is scarce. While shortage of individual-level data has not been of particular concern in job-seeker profiling, future work may explore the benefits of expert-augmentation in public processes where good quality data is especially hard to come by.

This thesis highlighted the importance of integrating public sector data-driven systems with qualitative domain knowledge and showed promising results in practice. To my best knowledge, it is the first study focused on improving machine learning models with expert knowledge specifically in the public sphere, not to mention the subfields of social and unemployment policy. This alone warrants that it contributes to academic research on public sector automated systems and knowledge engineering. Given this study confirmed advantages of the demonstrated approach from a purely technical perspective, follow-up research is welcomed to test whether it a) enhances justifiable and actionable decision-making from the perspectives of explainable AI and human-computer interaction and b) yields similar benefits for use cases with different policy- and performance priorities across the public sphere.

BIBLIOGRAPHY

Alexander, Patricia A (1992). "Domain knowledge: Evolving themes and emerging concerns". *Educational psychologist* 27 (1), 33–51.

Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager (2020). "Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective". *Frontiers in Big Data* 3.

Amershi, Saleema, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz (2019). "Guidelines for Human-AI Interaction". *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland UK: Association for Computing Machinery, 1–13.

Bailey, Diane E and Stephen R Barley (2020). "Beyond design and use: How scholars should study intelligent technologies". *Information and Organization* 30 (2), 100286.

Baldock, John, Sarah Vickerstaff, and Lavinia Mitton (2011). *Social policy*. Oxford University Press.

Baxter, Jonathan (2000). "A model of inductive bias learning". *Journal of artificial intelligence research* 12, 149–198.

Bekkers, Victor JJM and Stavros Zouridis (1999). "Electronic service delivery in public administration: Some trends and issues". *International review of administrative sciences* 65 (2), 183–195.

Bovens, Mark and Stavros Zouridis (2002). "From street-level to system-level bureaucracies: how information and communication technology is transforming administrative discretion and constitutional control". *Public administration review* 62 (2), 174–184.

Box, George EP (1976). "Science and statistics". *Journal of the American Statistical Association* 71 (356), 791–799.

Breiman, Leo (2001). "Random forests". *Machine learning* 45 (1), 5–32.

Cao, Li-Juan and Francis Eng Hock Tay (2003). "Support vector machine with adaptive parameters in financial time series forecasting". *IEEE Transactions on neural networks* 14 (6), 1506–1518.

Carey, Gemma and Mark Matthews (2017). "Methods for delivering complex social services: exploring adaptive management and regulation in the Australian National Disability Insurance Scheme". *Public Management Review* 19 (2), 194–211.

Coulet, Adrien, Malika Smaıl-Tabbone, Pascale Benlian, Amedeo Napoli, and Marie-Dominique Devignes (2008). "Ontology-guided data preparation for discovering genotype-phenotype relationships". *BMC bioinformatics* 9 (4), 1–9.

Dang, Hai-Anh H. and Cuong Viet Nguyen (2021). "Gender inequality during the COVID-19 pandemic: Income, expenditure, savings, and job loss". *World Development* 140, 105296.

Danziger, James N and Kim Viborg Andersen (2002). "The impacts of information technology on public administration: an analysis of empirical research from the "golden age" of transformation". *International Journal of Public Administration* 25 (5), 591–627.

Davies, Huw TO and Sandra M Nutley (2000). *What works?: Evidence-based policy and practice in public services*. Policy Press.

Deng, Changyu, Xunbi Ji, Colton Rainey, Jianyu Zhang, and Wei Lu (2020). "Integrating machine learning with human knowledge". *Iscience* 23 (11), 101656.

Desiere, Sam, Kristine Langenbucher, and Ludo Struyven (2018). *Profiling tools for early identification of jobseekers who need extra support. Policy Brief on Activation Policies*.

— (2019). *Statistical profiling in public employment services: An international comparison*.

Desiere, Sam and Ludo Struyven (2021). "Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off". *Journal of Social Policy* 50 (2), 367–385.

Dieterich, William, Christina Mendoza, and Tim Brennan (2016). "COMPAS risk scales: Demonstrating accuracy equity and predictive parity". *Northpointe Inc* 7 (4).

Dressel, Julia and Hany Farid (2018). "The accuracy, fairness, and limits of predicting recidivism". *Science advances* 4 (1), eaao5580.

Duan, Yanqing, John S. Edwards, and Yogesh K Dwivedi (2019). "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda". *International Journal of Information Management* 48, 63–71.

Dwivedi, Yogesh K., Laurie Hughes, Elvira Ismagilova, Gert Aarts, Crispin Coombs, Tom Crick, Yanqing Duan, Rohita Dwivedi, John Edwards, Aled Eirug, Vassilis Galanos, P. Vigneswara Ilavarasan, Marijn Janssen, Paul Jones, Arpan Kumar Kar, Hatice Kizgin, Bianca Kronemann, Banita Lal, Biagio Lucini, Rony Medaglia, Kenneth Le Meunier-FitzHugh, Leslie Caroline Le Meunier-FitzHugh, Santosh Misra, Emmanuel Mogaji, Sujeet Kumar Sharma, Jang Bahadur Singh, Vishnupriya Raghavan, Ramakrishnan Raman, Nripendra P. Rana, Spyridon Samothrakis, Jak Spencer, Kuttimani Tamilmani, Annie Tubadji, Paul Walton, and Michael D. Williams (2021). "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging

challenges, opportunities, and agenda for research, practice and policy". *International Journal of Information Management* 57, 101994.

Edwards, John S, Yanqing Duan, and PC Robins (2000). "An analysis of expert systems for business decision making at different levels and in different roles". *European Journal of Information Systems* 9 (1), 36–46.

Eurostat (2022). *Government expenditure on social protection*. Eurostat. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Government_expenditure_on_social_protection.

Flanagan, Mary, Daniel C Howe, and Helen Nissenbaum (2008). "Embodying values in technology: Theory and practice". *Information technology and moral philosophy* 322, 24.

Fokkema, Marjolein (2020). "Fitting Prediction Rule Ensembles with R Package pre". *Journal of Statistical Software* 92 (12), 1–30.

Friedman, Jerome H and Bogdan E Popescu (2008). "Predictive learning via rule ensembles". *The annals of applied statistics* 2 (3), 916–954.

Gennatas, Efstathios D, Jerome H Friedman, Lyle H Ungar, Romain Pirracchio, Eric Eaton, Lara G Reichmann, Yannet Interian, José Marcio Luna, Charles B Simone, Andrew Auerbach, et al. (2020). "Expert-augmented machine learning". *Proceedings of the National Academy of Sciences* 117 (9), 4571–4577.

Greenland, Sander, Judea Pearl, and James M Robins (1999). "Confounding and collapsibility in causal inference". *Statistical science* 14 (1), 29–46.

Höchtl, Johann, Peter Parycek, and Ralph Schöllhammer (2016). "Big data in the policy cycle: Policy decision making in the digital era". *Journal of Organizational Computing and Electronic Commerce* 26 (1-2), 147–169.

Huang, Jin and Charles X Ling (2005). "Using AUC and accuracy in evaluating learning algorithms". *IEEE Transactions on knowledge and Data Engineering* 17 (3), 299–310.

Hughes, Thomas P et al. (2012). "The evolution of large technological systems". *The social construction of technological systems: New directions in the sociology and history of technology*, 45–76.

Jordan, M. I. and T. M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects". *Science* 349 (6245), 255–260.

Kemshall, Hazel (2001). *Risk, social policy and welfare*. McGraw-Hill Education (UK).

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018). “Human decisions and machine predictions”. *The quarterly journal of economics* 133 (1), 237–293.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). “Inherent trade-offs in the fair determination of risk scores”. *arXiv preprint arXiv:1609.05807*.

Kristal, Tali and Meir Yaish (2020). “Does the coronavirus pandemic level the gender inequality curve? (It doesn’t)”. *Research in Social Stratification and Mobility* 68, 100520.

Lin, Jau-Huei and Peter J Haug (2006). “Data preparation framework for preprocessing clinical data in data mining”. *AMIA annual symposium proceedings*. Vol. 2006. American Medical Informatics Association, 489.

Mirchevska, Violeta, Mitja Luštrek, and Matjaž Gams (2014). “Combining domain knowledge and machine learning for robust fall detection”. *Expert Systems* 31 (2), 163–175.

Molina, Mario and Filiz Garip (2019). “Machine learning for sociology”. *Annual Review of Sociology* 45, 27–45.

Montgomery, Jacob M and Santiago Olivella (2018). “Tree-Based Models for Political Science Data”. *American Journal of Political Science* 62 (3), 729–744.

Mulligan, Deirdre K and Kenneth A Bamberger (2019). “Procurement as policy: Administrative process for machine learning”. *Berkeley Tech. LJ* 34, 773.

Provost, Foster and Ron Kohavi (1998). “Guest editors’ introduction: On applied research in machine learning”. *Machine learning* 30 (2), 127–132.

Radulescu, Carmen Valentina, Georgiana-Raluca Ladaru, Sorin Burlacu, Florentina Constantin, Corina Ioanas, and Ionut Laurentiu Petre (JAN 2021). “Impact of the COVID-19 Pandemic on the Romanian Labor Market”. *SUSTAINABILITY* 13 (1).

Rajagopalan, Balaji and Mark W Isken (2001). “Exploiting data preparation to enhance mining and knowledge discovery”. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 31 (4), 460–467.

Schapire, Robert E, Marie Rochery, Mazin Rahim, and Narendra Gupta (2002). “Incorporating prior knowledge into boosting”. *ICML*. Vol. 2, 538–545.

Sinha, Atish P and Huimin Zhao (2008). “Incorporating domain knowledge into data mining classifiers: An application in indirect lending”. *Decision Support Systems* 46 (1), 287–299.

Soibelman, Lucio and Hyunjoo Kim (2002). “Data preparation process for construction knowledge generation through knowledge discovery in databases”. *Journal of Computing in Civil Engineering* 16 (1), 39–48.

Speybroeck, Niko (2012). “Classification and regression trees”. *International journal of public health* 57 (1), 243–246.

Studer, Rudi, V Richard Benjamins, and Dieter Fensel (1998). “Knowledge engineering: principles and methods”. *Data & knowledge engineering* 25 (1-2), 161–197.

Töötukassa (n.d.). *Unemployment insurance benefit*. `https://www.tootukassa.ee/en/services/unemployment-insurance-benefit`. Accessed on 2022.04.20.

Troya, Íñigo Martınez de Rituerto de, Ruqian Chen, Laura O Moraes, Pranjal Bajaj, Jordan Kupersmith, Rayid Ghani, Nuno B Brás, and Leid Zejnilovic (2018). “Predicting, explaining, and understanding risk of long-term unemployment”. *32nd Conference on Neural Information Processing Systems*.

Viljanen, Markus and Tapio Pahikkala (2020). “Predicting unemployment with machine learning based on registry data”. *International Conference on Research Challenges in Information Science*. Springer, 352–368.

Vriens, Dirk and Jan Achterbergh (2015). “Tools for Supporting Responsible Decision-Making?” *Systems Research and Behavioral Science* 32 (3), 312–329.

Waelbers, Katinka (2011). *Doing Good with Technologies: Taking Responsibility for the Social Role of Emerging Technologies*. Vol. 4. Springer Science & Business Media.

Winner, Langdon (1980). “Do artifacts have politics?” *Daedalus*, 121–136.

Wirtz, Bernd W, Jan C Weyerer, and Carolin Geyer (2019). “Artificial intelligence and the public sector — applications and challenges”. *International Journal of Public Administration* 42 (7), 596–615.

Yang, Qian, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman (2020). “Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design”. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 1–13.

Yildiz, Mete (2007). “E-government research: Reviewing the literature, limitations, and ways forward”. *Government Information Quarterly* 24 (3), 646–665.

Yu, Ting, Tony Jan, Simeon Simoff, and John Debenham (Jan. 2007). “Incorporating Prior Domain Knowledge Into Inductive Machine Learning”.

APPENDIX 1. DATA TYPES AND VARIABLE TRANSFORMATIONS

Variable	Data type	Explanation	Applied transformation
EMPLOYED AFTER 180 DAYS	binary	Whether or not observation resumed work within 180 days from registering as unemployed	no transformation
EMPLOYMENT SPELLS	continuous	Number of employment spells in the last 3 years	no transformation
UNIQUE EMPLOYERS	continuous	Number of unique employers in the last 3 years	no transformation
LAST STATUS	nominal (3 levels)	Last status/activity before unemployment	Different types of employment contracts were aggregated into level "employed"; "In prison", "military conscription" and "unknown" were aggregated into level "other"; and "caretaker", "sick", and "studied" into one level.
REASON FOR ENDING LAST SPELL	nominal (5 levels)	Reason for ending last employment spell	13 specific reasons were divided into 4 categories depending on whether the reason was related to the employer (such as lay-off), employee (such as incompetence), mutual agreement or end of contract. An additional level was reserved for those who were not employed at all.
MONTHS WITH PAYMENT	nominal ordered (1-24)	Number of months with payment in the last 2 years	no transformation
TIME SINCE LAST SPELL	nominal ordered (8 levels)	Time passed since last employment spell	no transformation
LAST FIELD OF EMPLOYMENT	nominal (11 levels)	Field of last employment spell	A total of 53 specific fields were aggregated into 11 more general employment fields. For example, "forestry/fishing/hunting", "agriculture and veterinary" and "husbandry" were labelled as "agriculture".
COMPETITION FOR SUITABLE VACANCIES	continuous	Ratio of the total number of active vacancies and competing job-seekers for jobs that are suitable for a particular observation. In simpler terms, this ratio indicates competition for suitable/desired job vacancies, the smaller the more competition there is.	no transformation
AGE GROUP	nominal ordered (6 levels)	Age group	Numeric age variable was converged into 6 age groups as seen in appendix 2.
REGION	nominal ordered (5 levels)	Region of residence	15 Counties were divided into five regions: "North", "West", "Central", "North-East" and "South". Observations with unknown county of residence were excluded.
DURATION OF LAST SPELL	nominal ordered (6 levels)	Duration of last employment spell	no transformation
LEVEL OF EDUCATION	nominal ordered (4 levels)	Level of education	"No education" and "unknown" were combined into one category

WORK CAPACITY	nominal (3 levels)	Capacity to work	no transformation
UNEMPLOYMENT DAYS IN THE LAST 3 YEARS	continuous	Number of unemployment days in the last 3 years	no transformation
REGISTERED JOB-SEEKERS AT THE SAME TIME	nominal (3 levels)	The total number of people that registered as unemployed around the same time ($\pm$ 15 days) as the observation.	no transformation
FIELD OF EDUCATION	nominal (11 levels)	Field of education	no transformation
DURATION OF ASSIGNED UIB	continuous	Duration of assigned UIB in days	no transformation
SUITABLE VACANCIES	nominal ordered (7 levels)	Number of job vacancies in the desired ISCO field	no transformation
LAST SPELL ISCO	nominal (11 levels)	ISCO category (occupation group) of the last employment spell	Observations from group "0" (military) were excluded due to the marginal size of the group potentially allowing identification of certain individuals.
SHORT SPELLS IN THE LAST 3 YEARS	continuous	Number of employment spells in the last 3 years that were shorter than 90 days	no transformation
ESTONIAN LANGUAGE SKILLS	nominal ordered (8 levels)	Level of Estonian language skills, including "missing" and "unknown".	no transformation
ASSIGNED UNEMPLOYMENT ALLOWANCE	continuous	Duration of assigned unemployment allowance in days	no transformation
SHARE OF JOB-SEEKERS RESUMING WORK	continuous	Share of job-seekers that exited unemployment within the last 30 days	no transformation
ASSIGNED UIB DAILY RATE	continuous	Assigned UIB daily rate in Euros.	no transformation
PRIOR UNEMPLOYMENT SPELLS	nominal ordered (7 levels)	Number of previous unemployment spells in the last 3 years	no transformation
COMPUTER SKILLS	nominal (6 levels)	Self-reported computer skills	no transformation
HAS E-MAIL	binary	Whether observation has provided the unemployment agency with a functional e-mail address	no transformation
HAS DRIVER'S LICENSE	binary	Whether observation has a valid driver's license (Estonian B-category license for regular cars)	no transformation
IS BOARD MEMBER	binary	Whether observation is a board member of some company or organization	no transformation

RECEIVED WAGE SUBSIDY	binary	Whether observation received wage subsidy in the last 3 years	no transformation
CITIZENSHIP	nominal (4 levels)	Citizenship	Four categories for “Estonian”, “Russian”, “other” and “undetermined”. Observations with unknown citizenship status were excluded.
LAST EMPLOYMENT CONTRACT	nominal (4 levels)	Type of last employment spell	Eight specific types of employment were aggregated into 4 general levels as seen in appendix 2

APPENDIX 2. STRATIFIED TRAINING AND TEST SAMPLES

	Levels (nominal) or mean and median (continuous)	**Training sample**	**Test sample**	**Short-term future test sample**	**Mid-term future test sample**
Total observations		54027	109469	31515	117815
EMPLOYED AFTER 180 DAYS	0	27054 (50.07%)	54882 (50.13%)	15416 (48.92%)	61585 (52.27%)
	1	26973 (49.93%)	54587 (49.87%)	16099 (51.08%)	56230 (47.73%)
EMPLOYMENT SPELLS	mean	2.57	2.57	3.26	3.14
	median	2	2	2	2
UNIQUE EMPLOYERS	mean	2.10	2.11	2.56	2.44
	median	2	2	2	2
LAST STATUS	Carer/sick/studied	4801 (8.89%)	9693 (8.85%)	1458 (4.63%)	8490 (7.21%)
	Employed	42805 (79.23%)	86768 (79.26%)	27925 (88.61%)	98161 (83.32%)
	Other	6421 (11.88%)	13008 (11.88%)	2132 (6.77%)	11164 (9.46%)
REASON FOR ENDING LAST SPELL	Contract deadline	13608 (25.19%)	27173 (24.82%)	7346 (23.31%)	31515 (26.75%)
	Employee related	14517 (26.87%)	29686 (27.12%)	8337 (26.45%)	33212 (28.19%)
	Employer related	7870 (14.57%)	16241 (14.84%)	9399 (29.82%)	24728 (20.99%)
	Mutual agreement/other	15062 (27.88%)	30388 (27.76%)	5803 (18.41%)	24095 (20.45%)
	Not employed	2970 (5.50%)	5981 (5.46%)	630 (2.00%)	4265 (3.62%)
MONTHS WITH PAYMENT	mean	11.33	11.39	14.93	13.32
	median	11	11	18	15
TIME SINCE LAST SPELL	< 3 months	37701 (69.78%)	76504 (69.89%)	25633 (81.34%)	89961 (76.36%)
	3-6 months	3297 (6.10%)	6737 (6.15%)	1356 (4.30%)	5243 (4.45%)
	6-12 months	2973 (5.50%)	5960 (5.44%)	1511 (4.79%)	5412 (4.59%)
	1-2 years	2422 (4.48%)	4852 (4.43%)	1031 (3.27%)	5208 (4.42%)
	2-3 years	1252 (2.32%)	2488 (2.27%)	484 (1.54%)	2512 (2.13%)
	3-5 years	2450 (4.53%)	5004 (4.57%)	681 (2.16%)	4031 (3.42%)
	> 5 years	962 (1.78%)	1943 (1.77%)	189 (0.60%)	1183 (1.00%)
	Unknown/missing	2970 (5.50%)	5981 (5.46%)	630 (2.00%)	4265 (3.62%)

LAST FIELD OF EMPLOYMENT	Agriculture	1686 (3.12%)	3536 (3.23%)	440 (1.40%)	2742 (2.33%)
	Business service	7960 (14.73%)	17575 (16.05%)	6067 (19.25%)	21218 (18.01%)
	Construction	6685 (12.37%)	13741 (12.55%)	3438 (10.91%)	13943 (11.83%)
	Education	946 (1.75%)	1855 (1.69%)	463 (1.47%)	1673 (1.42%)
	Health & social	1311 (2.43%)	2816 (2.57%)	845 (2.68%)	3286 (2.79%)
	Industry	10282 (19.03%)	20433 (18.68%)	5280 (16.75%)	18742 (15.91%)
	Other	6249 (11.57%)	10869 (9.93%)	1906 (6.05%)	9142 (7.76%)
	Personal service	8975 (16.61%)	17738 (16.20%)	6712 (21.30%)	24090 (20.45%)
	Public sector	753 (1.39%)	1666 (1.52%)	427 (1.35%)	1559 (1.32%)
	Retail	6345 (11.74%)	13519 (12.35%)	4421 (14.03%)	16020 (13.60%)
	Transport	2835 (5.25%)	5721 (5.23%)	1516 (4.81%)	5400 (4.58%)
COMPETITION FOR SUITABLE VACANCIES	mean	0.14	0.14	0.06	0.07
	median	0.09	0.09	0.03	0.04
AGE GROUP	< 20	2020 (3.74%)	4025 (3.68%)	1018 (3.23%)	5488 (4.66%)
	20-30	16019 (29.65%)	32397 (29.59%)	9940 (31.54%)	35771 (30.36%)
	31-40	11886 (22.00%)	23726 (21.67%)	7461 (23.67%)	27673 (23.49%)
	41-50	10718 (19.84%)	22189 (20.27%)	6241 (19.80%)	22882 (19.42%)
	51-60	12242 (22.66%)	24692 (22.56%)	5856 (18.58%)	22246 (18.88%)
	> 60	1142 (2.11%)	2440 (2.23%)	999 (3.17%)	3755 (3.19%)
REGION	Central	4667 (8.64%)	9515 (8.69%)	2516 (7.98%)	9546 (8.10%)
	North	21160 (39.17%)	42816 (39.11%)	15910 (50.48%)	55700 (47.28%)
	North-East	9335 (17.28%)	18985 (17.34%)	3812 (12.10%)	15763 (13.38%)
	South	13220 (24.47%)	26676 (24.37%)	6295 (19.97%)	24939 (21.17%)
	West	5645 (10.45%)	11477 (10.48%)	2982 (9.46%)	11867 (10.07%)
DURATION OF LAST SPELL	< 3 months	17960 (33.24%)	36092 (32.97%)	8215 (26.07%)	37316 (31.67%)
	3-12 months	14437 (26.72%)	29037 (26.53%)	8996 (28.55%)	30993 (26.31%)
	1-3 years	8780 (16.25%)	18031 (16.47%)	6912 (21.93%)	22011 (18.68%)
	3-10 years	6836 (12.65%)	14126 (12.90%)	4877 (15.48%)	16584 (14.08%)
	> 10 years	3044 (5.63%)	6202 (5.67%)	1885 (5.98%)	6646 (5.64%)
	Unknown/missing	2970 (5.50%)	5981 (5.46%)	630 (2.00%)	4265 (3.62%)
LEVEL OF EDUCATION	None/unknown	811 (1.50%)	1546 (1.41%)	885 (2.81%)	2849 (2.42%)
	Primary level	12786 (23.67%)	25944 (23.70%)	6412 (20.35%)	26983 (22.90%)
	Secondary level	26447 (48.95%)	53449 (48.83%)	15747 (49.97%)	58527 (49.68%)
	Tertiary level	13983 (25.88%)	28530 (26.06%)	8471 (26.88%)	29456 (25.00%)
WORK CAPACITY	Has work capacity	44604 (82.56%)	90596 (82.76%)	27644 (87.72%)	101085 (85.80%)
	Partial work capacity	7261 (13.44%)	14423 (13.18%)	3461 (10.98%)	14444 (12.26%)
	No work capacity	2162 (4.00%)	4450 (4.07%)	410 (1.30%)	2286 (1.94%)
UNEMPLOYMENT DAYS IN THE LAST 3 YEARS	mean	124.15	122.75	82.05	103.38
	median	12	10	0	0
REGISTERED JOB-SEEKERS AT THE SAME TIME	mean	5941.65	5938.33	9879.16	8024.23
	median	5964	5959	9012	8004

FIELD OF EDUCATION	Business & law	3941 (7.29%)	8119 (7.42%)	2547 (8.08%)	8942 (7.59%)
	Education	1143 (2.12%)	2223 (2.03%)	648 (2.06%)	2400 (2.04%)
	Humanities	1353 (2.50%)	2878 (2.63%)	927 (2.94%)	3105 (2.64%)
	IT & communication	1009 (1.87%)	2067 (1.89%)	713 (2.26%)	2478 (2.10%)
	Natural sciences	810 (1.50%)	1505 (1.37%)	360 (1.14%)	1367 (1.16%)
	Agriculture	1673 (3.10%)	3452 (3.15%)	670 (2.13%)	2753 (2.34%)
	Social sciences	643 (1.19%)	1384 (1.26%)	375 (1.19%)	1327 (1.13%)
	Unknown/missing	25172 (46.59%)	50808 (46.41%)	14639 (46.5%)	56443 (47.91%)
	Service	6455 (11.95%)	12969 (11.85%)	3792 (12.03%)	13796 (11.71%)
	Health	767 (1.42%)	1690 (1.54%)	609 (1.93%)	2241 (1.90%)
	Industry & construction	11061 (20.47%)	22374 (20.44%)	6235 (19.78%)	22963 (19.49%)
DURATION OF ASSIGNED UIB	mean	73.53	74.04	118.63	95.71
	median	0	0	0	0
SUITABLE VACANCIES	0	3213 (5.95%)	6576 (6.01%)	2082 (6.61%)	17585 (14.93%)
	1	11502 (21.29%)	23011 (21.02%)	5242 (16.63%)	18262 (15.50%)
	2	11576 (21.43%)	23702 (21.65%)	5581 (17.71%)	19155 (16.26%)
	3	9168 (16.97%)	18662 (17.05%)	4602 (14.60%)	15490 (13.15%)
	4	6607 (12.23%)	12994 (11.87%)	3576 (11.35%)	12444 (10.56%)
	5	4175 (7.73%)	8683 (7.93%)	3079 (9.77%)	10475 (8.89%)
	6 or more	7786 (14.41%)	15841 (14.47%)	7353 (23.33%)	24404 (20.71%)
LAST SPELL ISCO	1	3439 (6.37%)	4637 (4.24%)	1681 (5.33%)	6719 (5.70%)
	2	2855 (5.28%)	7721 (7.05%)	2717 (8.62%)	8343 (7.08%)
	3	4039 (7.48%)	8615 (7.87%)	3079 (9.77%)	9998 (8.49%)
	4	2295 (4.25%)	5284 (4.83%)	1679 (5.33%)	5809 (4.93%)
	5	9365 (17.33%)	19695 (17.99%)	7214 (22.89%)	26312 (22.33%)
	6	763 (1.41%)	1537 (1.40%)	188 (0.60%)	927 (0.79%)
	7	10450 (19.34%)	21429 (19.58%)	5300 (16.82%)	18979 (16.11%)
	8	5828 (10.79%)	11687 (10.68%)	3137 (9.95%)	11367 (9.65%)
	9	11521 (21.32%)	21884 (19.99%)	5175 (16.42%)	23133 (19.64%)
	Missing	2970 (5.50%)	5981 (5.46%)	630 (2.00%)	4265 (3.62%)
	Unknown	502 (0.93%)	999 (0.91%)	715 (2.27%)	1963 (1.67%)
SHORT SPELLS IN THE LAST 3 YEARS	mean	1.19	1.19	1.49	1.52
	median	0	0	0	1
ESTONIAN LANGUAGE SKILLS	Missing	5674 (10.50%)	11336 (10.36%)	1979 (6.28%)	7783 (6.61%)
	Unknown	1476 (2.73%)	2844 (2.60%)	284 (0.90%)	968 (0.82%)
	A1	7936 (14.69%)	16070 (14.68%)	3148 (9.99%)	11859 (10.07%)
	A2	862 (1.60%)	1895 (1.73%)	1686 (5.35%)	6684 (5.67%)
	B1	839 (1.55%)	1778 (1.62%)	1745 (5.54%)	6835 (5.80%)
	B2	4494 (8.32%)	9013 (8.23%)	2289 (7.26%)	8120 (6.89%)
	C1	1795 (3.32%)	3632 (3.32%)	1942 (6.16%)	7642 (6.49%)
	C2	30951 (57.29%)	62901 (57.46%)	18442 (58.52%)	67924 (57.7%)
ASSIGNED UNEPLOYMENT ALLOWANCE	mean	60.63	61.36	52.71	60.95
	median	0	0	0	0

SHARE OF JOB-SEEKERS RESUMING WORK	mean	0.13	0.13	0.09	0.10
	median	0.12	0.13	0.11	0.11
ASSIGNED UIB DAILY RATE	mean	0.16	0.17	0.28	0.23
	median	0	0	0	0
PRIOR UNEMPLOYMENT SPELLS	0	26347 (48.77%)	53575 (48.94%)	18497 (58.69%)	62379 (52.95%)
	1	13548 (25.08%)	27726 (25.33%)	7295 (23.15%)	29156 (24.75%)
	2	7020 (12.99%)	13909 (12.71%)	2978 (9.45%)	13336 (11.32%)
	3	3775 (6.99%)	7476 (6.83%)	1438 (4.56%)	6769 (5.75%)
	4	1705 (3.16%)	3453 (3.15%)	642 (2.04%)	3111 (2.64%)
	5	811 (1.50%)	1756 (1.60%)	317 (1.01%)	1490 (1.26%)
	6 or more	821 (1.52%)	1574 (1.44%)	348 (1.10%)	1574 (1.34%)
COMPUTER SKILLS	None	197 (0.36%)	413 (0.38%)	605 (1.92%)	2855 (2.42%)
	Basic	4291 (7.94%)	8803 (8.04%)	7711 (24.47%)	31569 (26.80%)
	Adept	7391 (13.68%)	15176 (13.87%)	16171 (51.31%)	63618 (54.00%)
	Specialist	1541 (2.85%)	2997 (2.74%)	2963 (9.40%)	10516 (8.93%)
	Expert	306 (0.57%)	568 (0.52%)	743 (2.36%)	2236 (1.90%)
	Unknown	40301 (74.5%)	81512 (74.5%)	3322 (10.54%)	7021 (5.96%)
HAS E-MAIL	0	5247 (9.71%)	10109 (9.23%)	958 (3.04%)	3903 (3.31%)
	1	48780 (90.29%)	99360 (90.77%)	30557 (96.96%)	113912 (96.69%)
HAS DRIVER'S LICENSE	0	24750 (45.81%)	49954 (45.63%)	12807 (40.64%)	50412 (42.79%)
	1	29277 (54.19%)	59515 (54.37%)	18708 (59.36%)	67403 (57.21%)
IS BOARD MEMBER	0	53423 (98.88%)	108167 (98.81%)	30021 (95.26%)	113345 (96.21%)
	1	604 (1.12%)	1302 (1.19%)	1494 (4.74%)	4470 (3.79%)
RECEIVED WAGE SUBSIDY	0	52394 (96.98%)	105985 (96.82%)	30488 (96.74%)	113602 (96.42%)
	1	1633 (3.02%)	3484 (3.18%)	1027 (3.26%)	4213 (3.58%)
CITIZENSHIP	Estonian	43653 (80.80%)	88449 (80.80%)	26345 (83.60%)	98101 (83.27%)
	Russian	3899 (7.22%)	7914 (7.23%)	1863 (5.91%)	7351 (6.24%)
	Other	756 (1.40%)	1559 (1.42%)	700 (2.22%)	2454 (2.08%)
	Undetermined	5719 (10.59%)	11547 (10.55%)	2607 (8.27%)	9909 (8.41%)
LAST EMPLOYMENT CONTRACT	Employment contract	42583 (78.82%)	86326 (78.86%)	25997 (82.49%)	92892 (78.85%)
	Not employed	2970 (5.50%)	5981 (5.46%)	630 (2.00%)	4265 (3.62%)
	Obligation contract	7193 (13.31%)	14446 (13.20%)	4074 (12.93%)	16871 (14.32%)
	Other	1281 (2.37%)	2716 (2.48%)	814 (2.58%)	3787 (3.21%)

APPENDIX 3. 20 DECISION RULES SUBJECT TO EXPERT ASSESSMENT

ID	Subpopulation defined by decision rule	N cases	Coef.	Empirical risk	Avg. Exp. assessment	Expert STDV	Penalty	R
116	Time since last employment spell = unknown/no spell Competition for suitable job vacancies = 0.03 Age group = 51-60	5353	0.154	0.776	0.964	0.094	1.994	5
61	N of employment spells in the last 3 years = 1 N of months with payment in the last 2 years = 24 Time since last employment spell = up to 3 months N of unemployment days in the past 3 years = 1.1 Length of assigned UIB in days = 0	1714	-0.142	0.294	0.083	0.129	1.632	5
15	N of employment spells in the last 3 years = 0 Time since last employment spell = 3 to 5 years Age group = 51-60	4770	0.166	0.824	0.964	0.094	1.483	4
69	Time since last employment spell = up to 3 months Duration of last employment spell = 3 to 12 months Length of assigned UIB in days = 230.9 Received wage subsidy in the past 3 years = no	2806	-0.174	0.397	0.571	0.122	1.427	4
25	Time since last employment spell = 3 to 5 years Age group = 51-60	8853	0.215	0.754	0.929	0.189	0.924	3
109	N of employment spells in the last 3 years = 3.7 Time since last employment spell = up to 3 months Work capacity = has work capacity Has e-mail account = yes	29837	-0.140	0.403	0.250	0.204	0.752	2
105	N of employment spells in the last 3 years = 0.2 N of months with payment in the last 2 years = 0 Has e-mail account = no	1915	0.152	0.897	0.964	0.09	0.716	2
21	Time since last employment spell = missing/unknown Age group = 20-30 Has driver's license = no	3101	0.136	0.607	0.964	0.250	0.571	2
13	Time since last employment spell = missing/unknown Has e-mail account = no	2574	0.141	0.844	0.917	0.129	0.561	2
54	Time since last employment spell = up to 3 months Age group = less than 20 Has driver's license = no	344	-0.413	0.360	0.500	0.250	0.558	2
85	N of employment spells in the last 3 years = 1.9 Reason of ending last employment spell = employee-related N of months with payment in the last 2 years = 22.5 Time since last employment spell = up to 3 months Length of assigned UIB in days = 0	820	-0.641	0.09	0.179	0.189	0.467	2
36	N of employment spells in the last 3 years = 8.7 Time since last employment spell = up to 3 months	4314	-0.189	0.267	0.357	0.283	0.319	1

59	N of employment spells in the last 3 years = 4.2 N of months with payment in the last 2 years = 20.4 Length of assigned UIB in days = 9.9	7123	-0.166	0.237	0.179	0.189	0.311	1
42	N of employment spells in the last 3 years = 1.4 Time since last employment spell = up to 3 months Work capacity = partial work capacity Length of assigned UIB in days = 271.9	1501	0.204	0.736	0.786	0.225	0.220	1
137	Time since last employment spell = up to 3 months Work capacity = no work capacity	2055	0.429	0.834	0.786	0.267	0.181	1
97	N of employment spells in the last 3 years = 0.6 Work capacity = has work capacity N of unemployment days in the past 3 years = 288.4	5721	0.188	0.714	0.750	0.204	0.175	1
60	Reason of ending last employment spell = end of contract N of months with payment in the last 2 years = 15.8 Time since last employment spell = up to 3 months Length of assigned UIB in days = 23.4	4383	-0.314	0.215	0.250	0.204	0.172	1
41	N of employment spells in the last 3 years = 1 N of months with payment in the last 2 years = 22.2 Time since last employment spell = up to 3 months Age group = 20-30 Length of assigned UIB in days = 227.4	428	-0.138	0.449	0.429	0.278	0.072	1
126	Time since last employment spell = up to 3 months Age group = less than 20 Field of education = industry & construction	384	-0.136	0.435	0.429	0.189	0.033	1
57	N of employment spells in the last 3 years = 0.2 Time since last employment spell = unknown/missing Working capacity = no working capacity	1444	0.151	0.925	0.929	0.122	0.028	1

APPENDIX 4. EXPERT ASSESSMENT QUESTIONNAIRE GUIDE

Welcome! You are about to respond to a questionnaire for the MA thesis of Peeter Leets, a Master's student in the Johan Skytte Institute of Political Science. The title of the thesis is "AUGMENTING PUBLIC SECTOR DATA-DRIVEN DECISION SUPPORT SYSTEMS WITH EXPERT KNOWLEDGE: CASE OF OTT". In this survey, no personal data is collected, except for your e-mail address for validation purposes. There are a total of 20 questions and the survey is estimated to take about 20-30 minutes to complete.

Briefly about the study objective

The goal of my thesis is to improve upon Töötukassa machine learning model OTT by integrating qualitative expert judgement.

What is OTT?

OTT is a machine learning model that is actively employed in Töötukassa working processes. It predicts one's likelihood to resume work in 180 days after registering as unemployed based on a wide array of associated factors. These factors include general socio-demographic variables but also more specific labor market predictors such as assigned uneployment benefits and last employment field.

What is the objective of this survey?

The output of Machine Learning models tends to be overly probabilistic. Relationships modelled by machines -- why does the machine predict higher score to some person compared to another -- can often be incomprehensible to human interpreters. In order for the model to be more intuitive and accurate, it needs to be augmented with value judgement and experience of domain experts.

Example question and guide to answering

You will be presented (in random order) job-seeker subgroups that have been defined by a small number of factors. Based on these factors, you will be prompted to assess the probability of resuming work in 180 days after registering as unemployed -- of a person belonging to **that group** compared to the probability of the **average job-seeker**.
Hypothetical question and example of answering logic:

Factor 1
Number of months with payment in the last 2 years = **6.4** (range: 2-24)
Mean for all job-seekers = **11.8** (range: 0-24)

Factor 2
Working capacity = **partial working capacity**
Mode for all job-seekers = **has working capacity** (all levels: has working capacity, partial working capacity, no working capacity)

Compared to the average job-seeker, how would you assess the probability of resuming work in 180 days after registering as unemployed for a job-seeker belonging to the subgroup in question?

Possible choices: **significantly lower**, **lower**, **equal**, **higher**, **significantly higher**, cannot say

Explanation: The first factor is the number of months in the last 2 years before registering as unemployed. Since this is a numeric variable, group mean will be presented to you - on average, these job-seekers received salary on 6.4 months out of 24. The "range" in parentheses shows all values that were present in that group. In this case, there were people who received salary for 2 months, for 24 months, or somewhere in between these numbers. The blue figure below (11,8) represents the equivalent statistic for the entire job-seeker population interpretable as the "average" job-seeker. The second factor is work capacity. As it is a nominal variable, group mode (most frequently occurring value) will be displayed instead. We can see that group mostly consists of job-seekers with **partial working capacity**. Again, blue represents the mode for the entire job-seeker population and possible values for that variable. I ask you -- the expert -- to assess, how likely is someone from this subgroup to resume work in 180 days after registering as unemployed compared to the entire job-seeker population. **Since in this case, a person belonging to that hypothetical subpopulation 1) received payment less frequently in the past 24 months and 2) is likely only partially work-capable, one may assume their probability is lower or significantly lower than the probability for the entire population.**

NB

- It may seem like the sub-group defining variables have been selected by the algorithm rather randomly and there is not enough information to make an educated assessment. Regardless, I ask you to give your best guess based on your intuition. It is possible to respond with "cannot say" but I encourage you to use that option as seldom as possible -- even the general direction of your answer (lower or higher) is useful information for my analysis.
- Participation in this survey is voluntary and anonymous. By participating, you agree that your responses may be analysed for purpose of the aforementioned study and nothing else.
- Three randomly picked respondents will receive an Apollo gift card worth 10€, recipients will be contacted via e-mail by 1st of June 2020 at the latest.
- In case of questions and/or issues, contact Peeter Leets, the owner of the survey at [email] or [phone].