Predicting, explaining, and understanding risk of long-term unemployment

Íñigo Martínez de Rituerto de Troya
Universidade Nova de Lisboa

Ruqian Chen
University of Washington

Laura O. Moraes
Universidade Federal do Rio de Janeiro

Pranjal Bajaj
Columbia University

Jordan Kupersmith
University of California, Berkeley

Rayid Ghani
University of Chicago

Nuno B. Brás
Instituto de Telecomunicações

Leid Zejnilovic
Universidade Nova de Lisboa

Abstract

Predictive models are used by public employment services to forecast registrants' risk of becoming long-term unemployed, and to develop proactive, preventative interventions to help them develop their skills and find work. Traditionally, simple statistical models are used due to their interpretability. However, we found that even simple models may not be understandable by the case workers who use them, defeating the performance/complexity trade-off to the detriment of individuals at risk. Recent advances in model explainability yield the possibility of using more accurate predictive models while at the same time providing understandable and actionable explanations. We compare the predictive precision of simple and complex models, and use SHAP (SHapley Additive exPlanations) values to provide informative risk scores and actionable recourse to assist case workers in helping at-risk individuals to understand which factors contribute to their risk.

1 Introduction

Long-term unemployment (LTU) is defined by Eurostat as a consecutive period of 12 months or longer of being unemployed while actively seeking work. In 2018, Portugal has the 6th highest rate of long-term unemployment in Europe [1]. Women, older workers, and people with disabilities are disproportionately affected, leading to increasing social and income inequalities, resulting in what the OECD calls a "broken social elevator" [1].

LTU risk prediction and profiling systems [2] have been widely implemented, *e.g.* in the United Kingdom [3, 4], Ireland [5, 6], Germany [7], the Czech Republic [8], the United States [9], and elsewhere. Traditionally, logistic regression models have been used, owing to their simplicity and interpretability. However, simplistic interpretable models can lead to unintended consequences for individuals who are classified incorrectly. As such, a design trade-off which was originally intended to protect people may, in some cases, do more harm than good. Furthermore, despite being interpretable in theory, these models are not always presented in an understandable way to counselors, thus losing the major benefit of using these algorithms.[1] There is also a risk that counselors may place excessive

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

trust in the algorithm and not question its decisions, or not trust it at all and ignore it altogether [10, 11], rendering these systems ineffective.

Recent efforts in opening up "black box" algorithms [12, 13] now enable the use of more complex machine learning models in domains where accountability and transparency are critical to human well-being. In this paper, we explored the use of one such explainability framework, SHapley Additive exPlanations, which provides explanations for each individual about which factors contribute to increasing or decreasing their risk score – feedback which can lead to better personalised interventions.

In the EU, GDPR [14] introduced a "right to explanation" which requires automated decision systems to be made explainable to the people whom they make decisions about. However, the comprehensiveness of this mandate is contested [15] and some believe that it is insufficient for providing stakeholders with the necessary level of understanding to truly comprehend these systems' reasoning [16, 17, 18]. Nevertheless, tools for enlightening algorithmic decisions exist, such as the explanation framework used in the present work. Such explanations may help individuals understand what factors contribute to their LTU risk, and can thus be offered as constructive feedback to aid in remedying their situation ("gaming the system" [19]). Crucially, counterfactual explanations may provide individuals with actionable recourse for changing their prospects in light of their profiling [20, 21].

Specifically, within the context of this study, career counselors draft a *Personal Development Plan* (PDP) together with the registrant. The counselor aggregates feedback from conversations with the registrant as well as from their digital profile, which includes an LTU risk score. The LTU risk score is just one component of the overall decision making process, and does not automatically assign a PDP to an individual without the oversight of their counselor, though it may be used to suggest certain PDPs over others. As such, our proposal is a decision support tool (DST) within a human-in-the-loop (HITL) system, involving a high level of caseworker discretion [2].

2 Data

We obtained data from the Portuguese National Institute of Employment and Professional Development (IEFP), describing the professional backgrounds and sociodemographic profiles of 3.5 million people registered between 2007-2017, along with transactional records regarding their interactions with the institute (training courses undertaken, job interviews and job offers received, their attendance to such interventions, etc.). The internal features are listed on Table 1. At each interaction, an individual's *state* within the system is updated (possible *states* include: registered; changed category (changed from employed to unemployed or vice-versa); occupied; and unregistered.

In order to represent the economic backdrop against which these interactions took place, we also gathered publicly available macroeconomic indicators from various sources. Firstly, nationwide socioeconomic indicators (reported yearly, quarterly, or monthly) were obtained from PORDATA [22]. Additionally, biannual data at the level of municipalities were also obtained from PORDATA [23]. Finally, 2011 census data [24] were used to capture the differences across parishes.

Table 1: Internal features.

Demographic	Geographic	Professional	Transactional
Age	Region	Employment status	Registration duration (current)
Gender	Municipality	Current industry [25]	Registration duration (total)
Nationality	Neighborhood	Desired industry [25]	Motivation (proxies) [2]
Civil status	Urban/rural	Unemployment subsidy	Previously had a PDP
# Dependents		Social welfare	# interventions (last 1/3/5 years)
Education		Seeking part/full-time	Job offers received
Disability		Experience (time)	Job offers declined

[1] In interviews with career counselors, the authors learned that some counselors don't understand the current risk score report and thus ignore it altogether when developing an individual's Personal Development Plan.

[2] In interviews conducted by the authors, counselors described that a person's apparent interest or motivation in the process was a significant factor in their decision making process. Lacking this information, we approximated individuals' motivation by their unregistration motives.

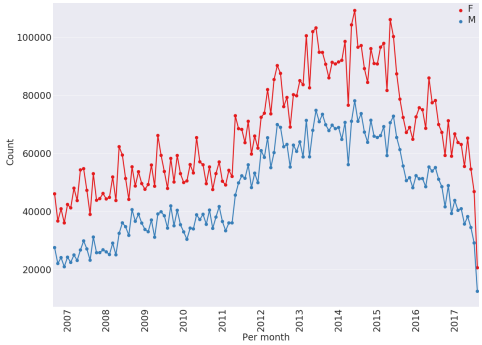

Figure 1: Unemployed registrants by gender.

Figure 2: Dashboard showing an individual's risk score, contributing factors, and risk score history.

2.1 Bias & fairness

There are intrinsic systemic biases reflected in the original data, particularly due to age, gender, and disability status. Long-term unemployment is especially prevalent among older workers, who find it difficult to reintegrate into the workforce after losing their job. This can adversely affect their pensions, and thus reduce their quality of life [26]. There are also consistently more women than men registered at the institute (figure 1), and proportionately more women were LTU than men at any point between 2007-2017. Individuals with disabilities are also at a much higher risk of being LTU.

Protected categories (age, gender, nationality, disability status, marital status)[27, 28] are kept in the model to allow for bias auditing and for future comparisons to human-only decisions. Specifically, the model should ideally achieve parity on both false negatives and false positives, across protected categories. False negative parity would ensure that all individuals who require additional assistance are not discriminated against by the system. False positive parity will ensure that no particular group of people is disproportionately led towards more intensive PDPs than they need.

3 Methods

3.1 Representation

Individuals were represented in monthly time-stamped snapshots describing their current and historical information at that point in time (*e.g.* current age, cumulative sum of past interventions, etc.) At each snapshot, the target label indicated whether or not the individual would be LTU 12 months in the future. A risk score is generated at the moment an individual first registers, or re-enters the system.

3.2 Temporal Cross-Validation, Evaluation, & Model Selection

In order to evaluate how well the model was able to predict future labels, we performed temporal cross-validation, splitting folds of $[t_0, t_0 + N_{\text{train}}]$ years for training, and $[(t_0 + N_{\text{train}}) + 1, (t_0 + N_{\text{train}}) + 1 + N_{\text{test}}]$ for testing, where t_0 is the first year of the training set, N_{train} and N_{test} are the number of years in the train and test sets respectively, and $+1$ ensures that the train set always precedes the test set. Temporal cross-validation was performed over the years 2007-2014 for model tuning, and finally, the models were retrained on the entire set $[2007, 2014]$ and tested on $[2015, 2016]$ for model evaluation. Data were pre-processed to remove new entries during the final year of the test sets (*i.e.* new registrations in that year; existing & continuing registrations were preserved) to account for the fact that an individual that enters in the final year of the set would not have a consecutive 12 month period in which to possibly become LTU.

Due to IEFP's limited resources, models were evaluated according to their precision at K people, with $K = 10\%$ (of the total population in the corresponding dataset), in order to ensure that the individuals with the highest risk score were most precisely profiled. This way, IEFP can ensure that

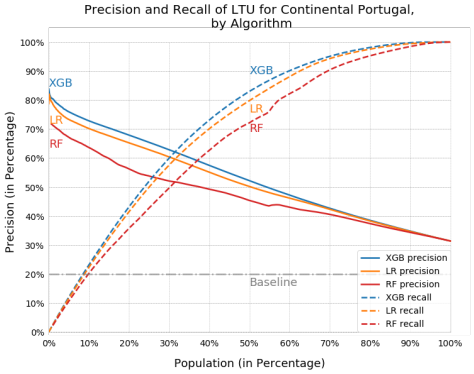

Figure 3: National model comparison.

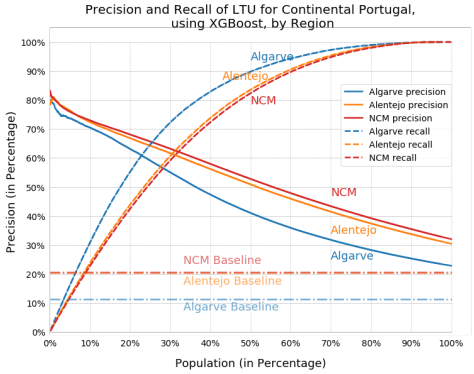

Figure 4: Regional model comparison.

their resources are going to help those who need them most. To ensure robustness across time, the final model was chosen to minimise the coefficient of variation, $\frac{\sigma}{|\mu|}$, of P@K, across temporal folds.

3.3 Model Explainability

SHAP values [12] are a form of additive feature attribution method which attribute the effect of individual features on a model's final output. Feature attributions for a *complex model* are found by building a simpler *explanation model* which approximates the behaviour of the original model. Feature attributions are then calculated by toggling features on/off, and observing their effects on the model's output. SHAP values explain a model's factors for *each* individual, rather than just providing an overall feature importance for the whole population. Furthermore, SHAP values are model-complete [29] and have been shown to be consistent with human intuition [12].

4 Results

Figure 3 shows the precision and recall curves for logistic regression, random forest and XGBoost models trained on the nationwide data. XGBoost outperforms logistic regression (the current state of the art in LTU risk profiling) consistently. The random forest model was outperformed by both. Figure 4 shows that training separate models stratified by region can achieve a similar precision curve but a sharper recall curve. (NB. NCM corresponds to a single regional aggregate model for the North, Center and Metropolitan Lisbon Area regions.) The baselines are the LTU rates of the datasets.

A dashboard (figure 2) shows an individual's current risk score along with its historical values. It also shows a personalised list of important contributing factors to LTU risk, provided by SHAP. Red indicates factor that increase risk while green indicates factors that decrease risk.

5 Conclusion

Algorithmic systems for decision making and decision support are increasingly being used in government and public institutions [30]. It is important that decisions made by these systems be interpretable by those who manage them (public servants, case workers, etc.) and that they provide actionable recourse for the individuals whom they are making those decisions about or for. While it is unrealistic to require public-facing civil servants to understand the inner-workings of a predictive algorithm, we should facilitate their understanding of these systems' ultimate decisions. Trust in these systems is also crucial if the users are to actually use system output, and not just ignore it.[1]

In this paper, we discussed how algorithmic decision support tools can be used to predict individuals' risk of becoming long-term unemployed while providing decision explanations that can be understood by a non-technical human agent. We showed that while simple models such as logistic regression can achieve reasonable precision, more complex models, such as XGBoost, yield improved performance. We used SHAP values to measure personalised feature attributions and designed an analytics dashboard to present the results in an understandable format for case workers (figure 2).

Acknowledgments

This work was supported in part by Nova School of Business and Economics and the Municipality of Cascais through the Data Science for Social Good Europe Fellowship, Fundação para a Ciência e a Tecnologia grant *"Avaliação de risco de desemprego de longa duração"*, AWS Cloud Credits for Research by Amazon.com Inc, and an Azure for Research Award by Microsoft Corporation. We would also like to thank our partners at IEFP for contributing their domain expertise, and the NIPS reviewers for their insightful feedback.

References

[1] *A Broken Social Elevator? How to Promote Social Mobility*. OECD Publishing, Paris, France, 2018.

[2] Annette Scoppetta and Arthur Buckenleib. Tackling long-term unemployment through risk profiling and outreach. a discussion paper from the employment thematic network. *European Commission - ESF Transnational Cooperation. Technical Dossier no. 6. Luxembourg: Publications Office of the European Union*, 2018.

[3] Simon Matty. Predicting likelihood of long-term unemployment: the development of a UK jobseekers' classification instrument. *Department of Work and Pensions, Working Paper No 116*, 2013.

[4] Clive Payne and Joan Payne. Early identification of the long-term unemployed. *Policy Studies Institute, Research Discussion Paper 4*.

[5] Philip J O'Connell, Seamus McGuiness, Elish Kelly, and John Walsh. National profiling of the unemployed in Ireland. *The Economic and Social Research Institute, Research Series*, (10), 2009.

[6] Philip J O'Connell, Seamus McGuiness, and Elish Kelly. A statistical profiling model of long-term unemployment risk in Ireland. *The Economic and Social Research Institute, Working Paper*, (345), 2010.

[7] Alexander Spermann. How to fight long-term unemployment: Lessons from germany. *IZA Institue for Labor Economics, Discussion Paper*, (9134), 2015.

[8] Tomáš Soukup. Profiling: Predicting long-term unemployment at the individual level. *Central European Journal of Public Policy*, 5(1), 2011.

[9] David Wiczer. Long-term unemployment: Attached and mismatched? *Federal Reserve Bank of St. Louis, Working Paper*, (2015-042A), 2015.

[10] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias - a hidden issue for clinical decision support system use. *Studies in Health Technology and Informatics*, 164:17–22, 2011.

[11] Angèle Christin. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, 4(2):1–14.

[12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[14] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[15] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2):76–99, 2017a.

[16] Lilian Edwards and Michael Veale. Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for. *16 Duke Law & Technology Review*, 18, 2017.

[17] Lilian Edwards and Michael Veale. Enslaving the algorithm: From a 'right to an explanation' to a 'right to better decisions'? *IEEE Security & Privacy*, 16(3):46–54, 2018.

[18] Michael Veale, Max van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, QC, Canada.

[19] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete Problems in AI Safety. *ArXiv e-prints*, June 2016.

[20] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2017.

[21] Alexander Spangher, Berk Ustun, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning*, 2018.

[22] PORDATA. Base de dados de portugal. `https://www.pordata.pt/Portugal`.

[23] PORDATA. Base de dados dos municípios. `https://www.pordata.pt/Municipios`.

[24] Instituto Nacional de Estadística. Censos 2011. `http://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=ine_censos_publicacoes`.

[25] *Classificação Portuguesa das Actividades Económicas. Rev. 3*. Insitituto Nacional de Estatística, Lisboa, Portugal, 2007.

[26] United States Government Accountability Office. *Unemployed Older Workers:Many Experience Challenges Regaining Employment and Face Reduced Retirement Security*. GAO-12-445, 2012.

[27] Parliament of the United Kingdom. *Equality Act 2010 c. 15*.

[28] *Civil Rights Act of 1964. Public Law 88–352, 78 Stat. 241, enacted July 2, 1964*.

[29] Leilani Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An approach to evaluating interpretability of machine learning. In *Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics*, 2018.

[30] Zeynep Engin and Philip Treleaven. Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *The Computer Journal*, 2018.