Reviewer's opinion on the doctoral dissertation titled "Machine learning supported survival analysis" authored by Róbert Csalódi

The PhD dissertation of Róbert Csalódi is built around the topic of survival analysis, where the author creatively combines the methodology of survival analysis with other data science, machine learning, and statistical methodologies. The thesis excellently demonstrates that the author navigates the topic with great confidence, creatively integrating and developing methodologies. It is also commendable that, alongside methodological developments, the dissertation includes many interesting case studies, each of which holds scientific significance and effectively illustrates the importance and timeliness of the topic.

The author presents the scientific theses in a dissertation consisting of 6 chapters and an appendix, totaling almost 150 pages. The first chapter introduces the topic, its motivation, describes the research questions, and outlines the thesis. The second chapter provides an overview of the most important concepts of survival analysis, which are frequently referenced throughout the dissertation. The third, fourth, and fifth chapters represent the substantive main chapters of the dissertation, where the author's independent work is presented. In each of these chapters, a new methodology is introduced, combining survival analysis with machine learning and data science methodologies.

The third chapter combines survival analysis with EM algorithm-based clustering technique to creatively address the problem of heterogeneous survival models by jointly clustering explanatory variables and survival time. The introduced algorithm's effectiveness is demonstrated through various case studies: Li-ion batteries, survival chances for patients with prostate cancer, and COVID-19 mortality rates of countries. The publicity of the theses of the third chapter is guaranteed by two journal articles, one published in the MDPI journal "Data" and the other in the multidisciplinary open-access journal "IEEE Access".

In the fourth chapter, the author combines survival analysis with frequent itemset-based association rule mining algorithm to creatively address the competing risks problem and identify triggering events, applying the methodology to course failure-based student dropout prediction. The publicity of the theses of the fourth chapter is ensured by an article published in the MDPI journal "Mathematics".

In the fifth chapter, the author ingeniously combines survival analysis with sequential pattern mining algorithm, introducing time-dependent support and confidence functions to creatively handle the temporal characteristics of association rules. The

Budapesti Műszaki és Gazdaságtudományi Egyetem
Természettudományi Kar
Matematika Intézet
Sztochasztika Tanszék

1111 Budapest, Egry József utca 1, H. ép. V. em. 7.
1521 Budapest, Pf. 91
Tel: 463-1101, Fax: 463-1677
stoch-adm@math.bme.hu, www.math.bme.hu/~sztoch

methodology is applied to the analysis of medical records to predict diseases based on previously known disorders. The publicity of the theses of the fifth chapter is ensured by two Else vier journal articles: one paper is published in the open-access multidisciplinary journal "MethodsX" and another paper is published in Knowledge-based Systems, a reputable interdisciplinary journal in the field of artificial intelligence.

The sixth chapter briefly concludes the dissertation.

The author formulates three theses related to the substantive chapters, each of which is further divided into additional subtheses, resulting in a total of three main theses and eight sub-theses. The author's independent scientific results are acknowledged through these theses, which have received adequate publicity, as evidenced by five journal articles. According to Google Scholar, these articles have received a total of 13 independent citations.

Although not directly related to the dissertation, it may be worth mentioning that the candidate has several other publications not directly related to the theses, some of which are highly cited. This commendably demonstrates that the candidate is an active researcher beyond his doctoral research program.

The bibliography contains 157 scholarly references. The majority of these are recent articles published in international journals, indicating that the author is familiar with the relevant literature and is well-versed in it. It also demonstrates the international context, significance, and timeliness of the topic.

Overall, the dissertation is well-structured, effectively linking the three substantive chapters, where survival analysis is integrated with another data science or machine learning procedure in each chapter, presenting a cohesive picture. It is also commendable that each chapter concludes with case studies.

The language of the dissertation is generally good. The figures and tables are well-designed, enhancing the readability of the work.

It is commendable that the author embeds and formalizes the introduced methods mathematically; these are, in most cases, sufficiently precise and aid in better understanding.

The thesis reflects the candidate's substantial effort. The ideas are innovative, the analyses are meticulously executed, the evaluations are expertly conducted, and the case studies are inherently interesting. **The remarkable results presented in the thesis are deemed suitable for the public defense of the dissertation. Therefore, I recommend the dissertation for acceptance**. Considering the scientifically and practically significant and novel findings, **recommending the conferment of the PhD title is anticipated**.

Budapesti Műszaki és Gazdaságtudományi Egyetem
Természettudományi Kar
Matematika Intézet
Sztochasztika Tanszék

1111 Budapest, Egry József utca 1, H. ép. V. em. 7.
1521 Budapest, Pf. 91
Tel: 463-1101, Fax: 463-1677
stoch-adm@math.bme.hu, www.math.bme.hu/~sztoch

MŰEGYETEM 1782

Below are some questions related to the dissertation:

1. In the third chapter, in the student dropout case study, how does treating the problem as a single outcome problem limit the usefulness of the model considering that "Failing to account for competing risks leads to biased estimates"?
2. In the Covid case study, how was it taken into account that different countries were affected by the pandemic at different times?
3. What are the advantages of the heterogeneous survival model over the homogeneous one? To what extent can the model obtained with the novel procedure introduced by the author be considered homogeneous?
4. In Chapter 3, the author emphasizes that the introduced method's important contribution is providing interpretable local models. Does the author plan to compare his solution with other interpretable machine learning methods (any supervised learning method, e.g., neural networks + interpretable layers such as SHAP)?
5. Could the method presented in the fourth chapter be extended to continuous time?
6. In the fourth and fifth chapters, association rules are evaluated using the confidence metric. However, the confidence metric has several limitations, and often the lift metric is preferable. Why was the lift metric not considered in these studies?

Budapest, July 2, 2024

Molontay Roland

Roland Molontay, PhD
Associate Professor, Deputy Director
Institute of Mathematics, Department of Stochastics
Budapest University of Technology and Economics

Budapesti Műszaki és Gazdaságtudományi Egyetem
Természettudományi Kar
Matematika Intézet
Sztochasztika Tanszék

1111 Budapest, Egry József utca 1. H. ép. V. em. 7.
1521 Budapest, Pf. 91
Tel: 463-1101, Fax: 463-1677
stoch-adm@math.bme.hu, www.math.bme.hu/~sztoch