Review of the Ph.D. thesis
"Regression Models for Software Project Effort Estimation"
by Huynh Thai Hoc

Huynh Thai Hoc 's dissertation deals with estimating the effort of software projects, which is a non-trivial task given the different applications (ranging from demanding scientific and technical computations to information systems with simpler operations but large amounts of data). However, these estimates are necessary not only for evaluating the complexity of the software work, but also for planning the necessary software development time, its price and the related financing of developers from the knowledge of the solution of "similar" tasks.

In this focus, the PhD student builds on the work of his supervisor Assoc. Prof. Zdenka Prokopová, Assoc. Prof. Petr Šilhavý and Assoc. Prof. Radek Šilhavý and the dissertations of earlier PhD students and it can be said that a school has already been established at FAI UTB.

The contribution of this work is the introduction of innovative transfer learning and dataset analysis techniques, implemented to improve the accuracy of effort estimation.

The author extended the tools of the function point analysis (FPA), identified relevant categorical factors (some of them are new and have not yet been described in the literature) that contribute to improved effort estimation, applied non-trivial methods of multiple linear regression (MLR, technique is employed for statistical analysis to establish the connection between a dependent and two or more independent variables), random forest (based on decision trees) deep learning based on multilayer perceptrons, deep learning with balanced datasets; ensemble techniques, transfer learning etc. These approaches were tailored to the domain under study, and experimentally validated on a broad class of large datasets (including from China), interpreting in detail the results obtained and possible limitations of the methods used, e.g., that FPA relies on fixed values of complexity weights.

It can be confirmed that the focus of the thesis is topical, the objectives of the thesis are challenging with a clearly formulated own contribution, and the thesis is **dissertable**.

The contribution of the thesis is also an extensive overview of the current state of research in all its aspects, which permeates the entire dissertation and not just the introductory parts, as is usually the case. The readability of the work is enhanced by conceptual frameworks and illustrative algorithmic sequences of calculation steps.

The dissertation includes the application of LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) techniques, which provide deeper insight into the black box of prediction models.

Another strength of the paper is that it does not limit itself to descriptive characteristics of the identified factors, but in the experimental part uses advanced statistical tools including hypothesis testing and Pearson correlation analysis.

A telling characteristic of the work's contribution is the author's excerpt:: "*The practical implications of this thesis carry considerable significance for the software industry. Predictive models might improve effort estimation accuracy by balancing datasets based on categorical variables or applying transfer learning based on pre-trained models.*"

The specification of further research is also valuable and can be an inspiration for other PhD students.

In terms of content and graphics, the work is at a very good level, as is the language level of the written text in English.

Formal comments:

- Chapter 4. Experiments doesn't actually present any experiments, its focus is on conceptual frameworks. The results of the experiments are only in Chapter 5.
- There is inconsistent terminology on page 21: Fig. 3-1: "Transactional functions", in the text "transaction functions".
- A large number of hard-to-remember abbreviations are used in the text (their list takes up two pages) and their meaning is given only once, and other occurrences often lack verbal context, which complicates the readability of the text.
- Some parts of the text do not follow the format of mathematical symbols, which should be in italics: page 36: "p independent variables" - "*p* independent variables"; "n records" - "*n* records"; similarly on page 37: M, N, m.
- Page 53: There should be a blank line below Figure 4-2.

Questions to the dissertant:

1. "If the inverse $(\boldsymbol{XTX})^{-1}$ exists":
   Is it possible that this matrix is singular and its inverse does not exist?
2. Page 39, Fig. 3-10, neural network with 4 layers.
   How to choose the number of layers?
   On page 61 you mention the choice of two hidden layers "*to make the model simple*". However, according to Kolmogorov, an ANN with only one hidden layer is sufficient to quantify any function.

Conclusion:

PhD student Huynh Thai Hoc's dissertation demonstrated the author's overview of software engineering and modelling tools in evaluating software project effort estimates, as well as the author's ability to creatively use existing approaches to formulate his own methodology and approaches, combining their advantages.

Huynh Thai Hoc has applied the results in 22 publications in international forums – 7 in journals (4 of them with impact factors), 14 in conferences and one as a book editor. Therefore,

I recommend

Huynh Thai Hoc's Ph.D. thesis to be accepted by the Committee to be presented and defended in the Engineering Informatics study branch

Brno, December 3, 2023

Prof. RNDr. Ing. Miloš Šeda, Ph.D.
Institute of Automation and Computer Science
Faculty of Mechanical Engineering
Brno University of Technology

Opponent: Assoc. Prof. Oldřich Trenz, Ph.D.
Department: Department of Informatics, Faculty of Business and Economics,
Mendel University in Brno
Office address: Zemědělská 1, 613 00 Brno
Contact: oldrich.trenz@mendelu.cz; tel.: 545 13 22 67

Opponent's Report – Doctoral Thesis

Author: **Huynh Thai Hoc**
Title of dissertation: **Regression Models for Software Project Effort Estimation**

Degree programme: Engineering Informatics
Degree course: Software Engineering
Supervisor: Assoc. Prof. Ing. Zdenka Prokopová, CSc.
Consulting Supervisor: Assoc. Prof. Ing. Petr Šilhavý, Ph.D.

General Assessment

The dissertation addresses the issue of estimating effort in software development, specifically the estimation of work involved in software project development. The dissertation estimates are important for the field of software development and have a significant impact on project planning and the efficient allocation of resources within a project.

The topic of the dissertation is very current and reflects contemporary trends in software engineering and project management. The use of advanced machine learning methods in the context of SDEE (Software Development Effort Estimation) is innovative and has significant potential for practical application.

The dissertation itself provides a new approach in the area of software development effort estimation by introducing innovative methods such as transfer learning and data set analysis to increase the accuracy of effort estimation, specifically within the extension of the Function Point Analysis (FPA). In addition, the dissertation examines various approaches to identifying factors in function point analysis and relevant categorical factors that contribute to improving effort estimation, including multiple linear regression, neural networks. The author conducted partial experiments to identify factors influencing effort estimation, leading to more accurate estimates compared to basic models for estimating software effort. The dissertation also describes the application of techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), which allow for a deeper insight into the structure of predictive models, i.e., how this prediction is constructed. It can be concluded that the author of the dissertation conducted research focused on evaluating the effectiveness of pre-trained models and proposed the use of deep learning methods in combination with strategies for balancing categorical variables. The goal was to improve the estimation of software effort. The results show that including relevant factors and the use of deep learning methods, as well as transfer learning techniques, improves the estimation of effort in software development. This can be used in software development (software projects) for better planning and management of these projects and ultimately to optimize overall costs.

The dissertation deals with the issue of estimating effort in software development, or estimating the effort required to develop software projects. These estimates are important for the field of software development and have a great influence on project planning and also on the efficiency of resource allocation within the project.

The topic of the dissertation is very current and reflects current trends in software engineering and project management. The use of advanced machine learning methods in the context of SDEE (Software Development Effort Estimation) is innovative and has significant potential for practice.

The dissertation provides a new approach in the area of effort estimation in software development by introducing innovative approaches such as transfer learning and data set analysis to increase the accuracy of effort estimation, namely within the extension of the function point analysis (FPA). In addition, different approaches to identify function point analysis factors and relevant categorical factors that contribute to the improvement of effort estimation, including multiple linear regression, neural networks, are explored in the present work. The author performed sub-experiments to identify factories affecting the effort estimate, leading to more accurate estimates compared to basic software effort estimation models. The dissertation also describes the application of LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive Explanations) techniques, which allow a deeper insight into the form of prediction models, i.e., how this prediction is built. It can be stated that the author of the dissertation conducted research aimed at evaluating the effectiveness of pre-trained models and proposed the use of deep learning methods in combination with strategies for balancing categorical variables. The goal was to improve software effort estimation. The results show that the inclusion of relevant factors and the use of deep learning methods, as well as transfer learning techniques, improve software development effort estimation. This can be used in the development of software (software projects) for better planning and management of these projects and thus as a result to optimize overall costs.

Dissertation description

The dissertation is divided into 8 chapters. In the first chapter (Introduction), the author specifies the motivation, research questions (hypotheses) and the goal of the dissertation. This is followed by the theoretical part (chapter 2), which maps the methods and approaches available for solving the topic of the dissertation. The third chapter (Methodology), where the methodological procedure of the solution is conceived in a wider context, including a detailed analysis. This is followed by the fourth chapter devoted to own experiments (Experiments). These liaison chapters are followed by sections dedicated to analysis of results (chapter 5 - Results and Discussion), analysis and applicability of results (chapter 6 - Contributions), Threat and Validation and Conclusion.

The actual structure of the dissertation corresponds to the type of this work, i.e., a dissertation. The aim of the dissertation is clearly defined, while it is divided into several partial parts. Research questions (hypotheses) are also defined in the introductory chapter, but here I would assume that these hypotheses will be formulated only after the research that maps the current state of solutions in the given area, i.e. after the identification of gaps in current research. The theoretical part of the dissertation is covered in relatively detail, including a reference to the sources used. For some parts of the text, newer sources can also be found, confirming the facts formulated here, but this is not a fundamental error. In the chapter devoted to the methodology of the work, both the methodological procedure and the detailed research are outlined, and the theory is also analysed. From my point of view, the theoretical framework (its extensiveness, often even redundancy) reduces the clarity of this chapter, but everything essential is presented here. A key part is the chapter devoted to experiments. Here, the individual proposed methods (listed in the theoretical part and methodology of the disertation) and their modifications are tested. This part is evaluated in the following chapter (Results and discussion), so it is necessary to evaluate both parts at the same time. The chapter dedicated to the results is quite extensive, individual comparisons are made here, both in the form of published tables and graphs. Some parts are discussed in detail, but elsewhere the description is slightly limited. Selected hypotheses are also evaluated here (chapter 5.3). Overall, the last chapters (chapters 5, 6, 7 and 8) tell about the results of the work, their applicability and validity, together with a discussion, although it can be said that with regard to clarity, these chapters could be better designed.

The formal side of the work can be assessed as good, some graphic objects could be published in higher quality or in higher resolution (for example, Figure 3-3). Overall, I evaluate the output (dissertation theses) positively, it is a comprehensive approach to the issue of optimizing the estimation of software complexity of software product development.

The author's publishing activity is quite above standard (conferences 14, journals 7). It is necessary to mention that the author of the dissertation thesis has also published the issues elaborated within the dissertation thesis in journals and conferences including IEEE and Springer.

Positive aspects:

- The dissertation is well structured and contains a detailed analysis.
- The experiments are carefully designed and provide new insights into the effort estimation problem.
- The use of methods such as LIME and SHAP for the interpretation of models is beneficial for improving the transparency and comprehensibility of the results. Incorporation of soft-computing methods into the solution.
- Publication outputs of the author, i.e., publication of dissertation thesis outputs in selected journals.

Negative aspects:

- The dissertation could further develop the comparison with existing methods and provide a more detailed discussion of the limitations of the approaches used (limitations for their deployment).
- It would be appropriate to verify the applicability of the conclusions on a real project (case study).
- Some parts of the dissertation could be conceived more compactly and thus more clearly.
- The dissertation do not publish the assessed data (datasets) to a sufficient extent (software projects), including the assessed parameters

Questions and Recommendations

- What specific technical or methodological obstacles might you expect to encounter in transferring your models from theoretical research to real-world software engineering? How can these challenges be addressed?
- How do your results differ from traditional software engineering effort estimation methods? What changes would need to be made to existing processes (software development) to incorporate your findings?
- What are the key steps for incorporating your methods into a software development workflow? How would you ensure that your methods are flexible enough to accommodate different types and sizes of software projects?

- You use 5-fold cross-validation in your dissertation, does this setting of the approach (5-fold) give the best results, or have you tried other variants?

Overall Assessment

The dissertation is a high-quality contribution to the area of software development effort estimation, or software project development effort estimation. The author demonstrates a good understanding of the topic. Overall, it can be stated that Huynh Thai Hoc dissertation represents an interesting contribution to the field of estimating the effort required for software development, although certain aspects could be further elaborated and deepened (deploy ability conditions) already in the basic text.

In my opinion, the student showed his in-depth knowledge of the researched field and his ability of systemic work on the research topic. I conclude that the dissertation "Regression Models for Software Project Effort Estimation" meets all requirements set for the appropriate level of study and recommend it for defence. After an eventual successful defence, I recommend Huynh Thai Hoc to be awarded the doctoral degree "Ph.D.".

In Brno, 14. 11. 2023

Assoc. Prof. Oldřich Trenz, Ph.D.
Opponent's signature

Reviewer:
doc. Ing. Radek Matušů, Ph.D.
Faculty of Applied Informatics
Tomas Bata University in Zlín

The Review of the Doctoral Thesis Entitled **"Regression Models for Software Project Effort Estimation"** by *Huynh Thai Hoc*

The structure of this review is in accordance with Article 52 of The Study and Examination Regulations of Tomas Bata University in Zlín.

a) Thesis Topicality

The doctoral thesis deals with an important and present problem of software development effort estimation, which has a direct impact on the accuracy of planning and the precision of resource allocation. The necessity of the investigation, as well as the motivation of the student for his work in the field and for the elaboration of the thesis, are well described.

b) Fulfillment of the Thesis Aims

The reasonable objectives of the thesis are clearly stated in Section 1.4 by means of four points and supported by five research questions and hypotheses introduced in Section 1.3. In my opinion, all objectives were fulfilled completely and they are adequately discussed in Sections 5, 6, and 8. However, it would be useful to provide also a lucid summary with direct answers to the research questions raised in Section 1.3 and concluding comments related directly to the research objectives as defined in Section 1.4.

c) The Approach to Solving the Problem and the Doctoral Thesis Results

The student started naturally with state-of-the-art in the field of software development effort estimation. This part is well elaborated and provides a thorough and logically structured, about 15 pages long, literature review on the related problems and approaches. The methodology itself covers the important aspects of function point analysis and its extensions, data collection, preprocessing techniques, model development based on various selected approaches, model interpretability, and comparison criteria.

The student performed a large array of experiments. First, he provided the experimental framework and then he described the details of the applied techniques, including Multiple Linear Regression, Random Forest, Multilayer Perceptron, Transfer Learning, etc. I appreciate the large amounts of experiments on performance comparisons accompanied by the rich discussions. The comprehensive evaluations of effort estimation methods obtained represent the key results of the doctoral thesis.

d) Significance for Science and Practice

In my opinion, the obtained results significantly contribute to current research in software development effort estimation. The student not only highlighted the importance and meaning of his own findings but also identified possible future directions for research. From the practical viewpoint, the thesis results are potentially meaningful for the software industry without any doubts.

e) Formal Level and Linguistic Elaboration of the Thesis

The formal level of the doctoral thesis is high. We could discuss a few minor imperfections, but I cannot see any serious formal flaws. The structure is logical and balanced. The total extent of the thesis is 116 pages, which is adequate. The number of figures, tables, and references is also reasonable and makes sense with respect to the content of the thesis. In my opinion, the level of English in the thesis is very good.

f) Publication Activities of the Doctoral Student

The publication activities of the doctoral student are excellent. He is (co-)author of 7 journal papers (including 1 accepted) and 14 conference contributions, of which he is the first author of 4 journal papers and 4 conference contributions. I would like to highlight his considerable journal publications as the first author in 3 SCIE journals and 1 ESCI journal, where especially IEEE Access, despite being purely Open Access, represents a highly reputed journal in Q2 according to M17+ Methodology. Moreover, he participated in editing a book.

The majority of student's publication activities are directly related to the topic of the doctoral thesis. It proves that the results have already been accepted by the research community.

Questions for the Thesis Defense

Q1: How do you interpret the results of LIME as presented in Fig. 5-41 on page 90? Can the variables in the negative section of the graph be eliminated from the predictive model?

Q2: One of the selected comparison criteria is standardized accuracy (SA), which is defined in (15). As you are aware, the higher the SA value, the better the estimation method. Nevertheless, the obtained SA values are 0 in many of your results (Figs. 5-11, 5-12, 5-13, 5-14, 5-15, 5-16, 5-17, 5-34, and 5-37). How do you interpret these zero SA values?

Overall Evaluation

In my opinion, the reviewed doctoral thesis fulfills all requirements posed on the thesis aimed at obtaining a doctoral degree (Ph.D.). Thus, I **recommend** the doctoral thesis for the defense in front of the respective committee.

In Zlín, 24 November 2023

doc. Ing. Radek Matušů, Ph.D.