Report on Master Thesis

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

Student:	**Michal Todt**
Advisor:	**Petr Polák**
Title of the thesis:	**Estimating performance of classifiers from dataset properties**

OVERALL ASSESSMENT *(provided in English, Czech, or Slovak):*

Contribution

With the onset of big data, and data science in economics, machine learning based classification became popular topic since it can potentially help in many economic problems (nicely reviewed in the second Chapter of the thesis). The thesis contributes to the topic by exploring performance of several classification schemes under different distributional assumptions. On an Australian credit data, the thesis then utilizes results, and even thought the empirical application is not successful, the simulation results are interesting, and comparison of the modern classification methodologies is itself an important contribution.

Methods

Author uses advanced techniques requiring broad knowledge beyond the scope of the Master level curriculum at IES. The methods are well mastered, author provides codes to the exercise, and makes the results transparent.

Literature

Author shows deep understanding of the literature on classification, mainly provides nicely written discussion on classification in economics and finance.

Manuscript form

Although the text has good structure and form in general, and the text is logically organized, the presentation sometimes lacks clarity. In my view motivation could be improved by motivating the classification via classical economic examples first for example. Author rather opens the introduction with difficult to read discussion about the „trend“ of black-box interpretation, and leaves reader with further references and many open questions. Classification problems can rather be thought of as a probabilistic regression, i.e. classical probit/logit where probability of an event is modelled by a more general, possibly non-linear function. From a point of view of a statistician, machine learning based classification is just a semiparametric nonlinear probabilistic (or limited dependent variable if you like) regression, and black-box label is not so important since it is only due to the fact that we are not really interested in interpreting parameters (or weights). From a point of view of an economist, it can help in any field using the classical probabilistic models. This is discussed later, but reader would benefit from learning about the motivation earlier. Similarly to the motivation, presentation of results is not clear, and one quickly loses the insight on what is actually done, mixing simulations with data comparison, etc.

Report on Master Thesis

Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague

Student:	**Michal Todt**
Advisor:	**Petr Polák**
Title of the thesis:	**Estimating performance of classifiers from dataset properties**

Summary and suggested questions for the discussion during the defense

In conclusion, the thesis is solid piece of work which deserves to be defended. Although the author does not show usefulness of modern statistical classification techniques for a real economic problem, comparison and discussion on artificial data is satisfactory and interesting enough.

My main questions for the defense are:

1/ What is the main reason of the unsuccessful data matching
2/ Which areas of economics/finance can be (possibly) improved by studied classification techniques
3/ Are there any caveats in implementation of the techniques in practice, especially with big data?

In case author answers the questions confidently, I suggest the thesis to be defended with grade "B".

SUMMARY OF POINTS AWARDED *(for details, see below):*

CATEGORY	**POINTS**
Contribution (max. 30 points)	24
Methods (max. 30 points)	26
Literature (max. 20 points)	20
Manuscript Form (max. 20 points)	13
TOTAL POINTS *(max. 100 points)*	**83**
GRADE (A – B – C – D – E – F)	**B**

NAME OF THE REFEREE: Jozef Barunik

DATE OF EVALUATION: September 10, 2018

Digitally signed by Jozef Barunik 100000083222
DN: dc=org, dc=terena, dc=tcs, c=CZ, o=Charles University, cn=Jozef Barunik 100000083222
Date: 2018.09.11 21:26:10 +02'00'

Referee Signature

EXPLANATION OF CATEGORIES AND SCALE:

CONTRIBUTION: *The author presents original ideas on the topic demonstrating critical thinking and ability to draw conclusions based on the knowledge of relevant theory and empirics. There is a distinct value added of the thesis.*

Strong	*Average*	*Weak*
30	*15*	*0*

METHODS: *The tools used are relevant to the research question being investigated, and adequate to the author's level of studies. The thesis topic is comprehensively analyzed.*

Strong	*Average*	*Weak*
30	*15*	*0*

LITERATURE REVIEW: *The thesis demonstrates author's full understanding and command of recent literature. The author quotes relevant literature in a proper way.*

Strong	*Average*	*Weak*
20	*10*	*0*

MANUSCRIPT FORM: *The thesis is well structured. The student uses appropriate language and style, including academic format for graphs and tables. The text effectively refers to graphs and tables and disposes with a complete bibliography.*

Strong	*Average*	*Weak*
20	*10*	*0*

Overall grading:

TOTAL	**GRADE**
91 – 100	**A**
81 - 90	**B**
71 - 80	**C**
61 – 70	**D**
51 – 60	**E**
0 – 50	**F**