Report on **Master Thesis**

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

Student:	**Bc. Barbora Bajgarová**
Advisor:	**doc. PhDr. Jozef Baruník, Ph.D.**
Title of the thesis:	**Machine Learning Methods in Motor Insurance Fraud Detection**

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

Short summary

The thesis focuses on fraud detection in the insurance industry, which is an interesting problem not only because of the size of the industry. Furthermore, the econometric part of the problem deals with unbalanced data, which are often present in real world examples and need to be handled with care. The author does a good job of comparing different machine learning methods of approaching the given problem.

Contribution

Firstly, the main contribution of the thesis is a good solution and approach to a real-world problem that requires care in the details. Secondly, the author contributes to the literature with the comparison of relevant methods, such as logistic regression, random forest and gradient boosting techniques to the fraud detection problem/classification. As mentioned, the data requires the use of additional algorithms to handle unbalanced data sets, which is done with SMOTE and effects weighting. In addition, working with the unique data set (UNIQA pojišťovna), which required proper preprocessing and taking into account its peculiarities, such as hidden biases or the distribution problem of cases, is part of the contribution.

Methods

As the title suggests, the text deals with machine learning methods, particularly those used for classification. Although the range of methods is obviously wider than presented here, the author does a good job of selecting and using appropriate and diverse tools for the problem. The author focuses on supervised machine learning methods, which can be further divided into linear and non-linear, and parametric and non-parametric. A simple benchmark approach is logistic regression. The cornerstone of the thesis is the comparison of methods that the author has learnt herself and that are not part of the IES curriculum, with the exception of multilayer perceptrons. Such tools are Decision Trees, Random Forests, Gradient Boosting methods - XGBoost and Histogram-Based Gradient Boosting, and the already mentioned Multilayer Perceptrons. All of these methods are well established in the literature and appropriate to the problem. A well-known drawback of machine learning methods is the interpretability of the models and their results. The author recognises this problem and uses Shapley values for feature importance to interpret the impact of the data on the prediction task.

Literature

The author employs and references pertinent literature on the subject, given the vast and continually expanding body of literature on the topic. She incorporates relevant sources in the bibliography that are correctly cited in the text.

Manuscript form

The manuscript has correct form and fulfills the IES standards. The thesis is well written and unwinds the story of the problem, which is easy for reader to follow. It is well structured, all tables and figures are at correct place and clearly presented.

Report on **Master Thesis**

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

Student:	**Bc. Barbora Bajgarová**
Advisor:	**doc. PhDr. Jozef Baruník, Ph.D.**
Title of the thesis:	**Machine Learning Methods in Motor Insurance Fraud Detection**

Overall evaluation and suggested questions for the discussion during the defense

In general, the thesis meets the standards required for a master's thesis at the IES, Faculty of Social Sciences, Charles University. In my assessment, it is a well-executed work, incorporating a genuine world data illustration. I recommend the thesis for defense and propose a grade A.

The results of the Turnitin analysis do not indicate significant text similarity with other available sources.

Q1: Citing author's last sentence in performance comparison section: *„This can further help insurance companies to efficiently allocate resources available for claim investigation."* How does your results translate to real numbers in company cost saving given or not your UNIQA case?

Q2: How would you sum up the characteristics of fraud? Firstly, from the real data and secondly, from the prediction models.

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

CATEGORY		**POINTS**
Contribution	*(max. 30 points)*	28
Methods	*(max. 30 points)*	30
Literature	*(max. 20 points)*	18
Manuscript Form	*(max. 20 points)*	19
TOTAL POINTS	*(max. 100 points)*	**95**
GRADE	**(A – B – C – D – E – F)**	**A**

NAME OF THE REFEREE: Luboš Hanus

DATE OF EVALUATION: 16. 8. 2024

Digitálně podepsáno (16. 8. 2024):
Luboš Hanus

_____
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.*

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.*

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

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.*

Overall grading:

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