Report on **Bachelor / Master Thesis**

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

Student:	**Andrej Zidek**
Advisor:	**Prof. Ing. Janda Karel, Dr., Ph.D., M.A.**
Title of the thesis:	**Application of the logit leaf algorithm for customer churn prediction in the energy distribution industry in the Czech Republic**

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

Please provide a short summary of the thesis, your assessment of each of the four key categories, and an overall evaluation and suggested questions for the discussion. The minimum length of the report is 300 words.

Short summary
The author investigates the determinants of customer churn using data obtained from MND Energie, a.s., one of the largest Czech energy suppliers. It partially focused on the influence of a treatment - the notification of the possibility of a new fixed-price contract - on churn. The author used a dataset of about 9000 customers whose natural gas contracts terminated at the end of 2019 including various information -such as average consumption, tariff, and sociodemographic characteristics. After applying the logit leaf model (LLM) and its individual building blocks individually -namely, decision tree and logit regression, the author concluded that LLM has better predictive performance on churn than its building blocks.

Contribution
The approach is original in the sense that the author applied a relatively recent and relevant supervised learning algorithm on the topic. The author contributes to the customer churn litterature by providing an additional empirical study using original techniques and relatively sufficient data.

Methods
The author uses applicable „why not“ empirical techniques to the research question while not being taught in classical econometrics courses. However and in my view, there are some misunderstandings , at least as far as I'm concerned, for the following reasons. **1.** I have never seen a 5x2 fold cross-validation used to check robustness; **2.** In the Prediction methods section, under the Logistic regression subsection, not all the assumptions are written; **3.** In the same subsection, the author seems to be mixing up the terms "independent variable" and "dependent variable". **4.** In the same subsection, the author assumes independence of the independent variables which is strong, use ‚not correlated‘ instead; **5.** the author states that the MLE estimator is consistent, asymptotically normal and efficient without mentionning the underlying assumptions (appropriate chosen distribution for the error term, the error terms are independent between each other, the optimizer found the global maximum of the log-Likelihood function); **6.** the author states that the logistic regression is unsuitable when the target variable is dependent on interactions between the independent variables, can we add an interaction term in the model equation?; **7.** the meaning and use of the „No. of cons points“ independent variable is unclear and there might be a significant problem in the results due to it; **8.** The interpretation of the beta coefficients of the logistic regression is sometimes wrong; **9.** It is unclear how the prediction performances metrics was computed; **10.** finally the data is poorly described: no extensive descriptive statistics and boxplots provided, essential when talking about outliers in the data pre-processing section.

Report on **Bachelor / Master Thesis**

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

Student:	**Andrej Zidek**
Advisor:	**Prof. Ing. Janda Karel, Dr., Ph.D., M.A.**
Title of the thesis:	**Application of the logit leaf algorithm for customer churn prediction in the energy distribution industry in the Czech Republic**

Literature
The author seems to understand the key ideas of the cited literature. The quoted literature is often relevant. However, there is a very significant lack of citations within the text (missing source of the information stated in the text) and the quoting format is heterogeneous.

Manuscript form
Although the author has used appropriate language and style, the graphs and tables are poorly formatted (screenshots of R outputs) and the thesis is poorly structured. The sections are not numbered.

Overall evaluation and suggested questions for the discussion during the defense
In my view, the thesis fulfills the requirements for a bachelor thesis at IES, Faculty of Social Sciences, Charles University, I recommend it for the defense and suggest a grade C. The results of the Turnitin analysis do not indicate significant text similarity with other available sources.

Could the student clearly explain how the predictive performance metrics are computed (e.g., with a pseudo-code)?
Could the student explain and describe the meaning of the „No of cons points“ variable? Why is this variable important in the decision tree root node split? The parallel drawn between this variable and gender is questionable. What about excluding the „No of cons points“ variable and see if the decision tree chooses the gender variable for the root node split?
Could the student interprete the beta coefficient associated with the male category of the gender variable in the logistic regression results?

***SUMMARY OF POINTS AWARDED**:*

CATEGORY		**POINTS**
Contribution	*(max. 30 points)*	27
Methods	*(max. 30 points)*	19
Literature	*(max. 20 points)*	13
Manuscript Form	*(max. 20 points)*	12
TOTAL POINTS	*(max. 100 points)*	**71**
GRADE	**(A – B – C – D – E – F)**	**C**

NAME OF THE REFEREE: MATHIEU PETIT

DATE OF EVALUATION: 2023-08-19

Digitally signed (19. 8. 2023):
Mathieu Petit

_____

Report on **Bachelor / Master Thesis**

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

Student:	**Andrej Zidek**
Advisor:	**Prof. Ing. Janda Karel, Dr., Ph.D., M.A.**
Title of the thesis:	**Application of the logit leaf algorithm for customer churn prediction in the energy distribution industry in the Czech Republic**

Referee Signature