

MASTER STATISTICS & DATA SCIENCE. PROPOSAL FOR THE THESIS PROJECT

Period of the project

Carrying out the project will account for 30 ECTS, which is equivalent to 21 weeks of 40 hours.

- Intended period of the project: 01 / 11 / 2022 - 30 / 04 / 2023
- Intended date for the Midterm Progress Meeting: 31 / 01 / 2023

Student

Name: Felix Kapulla

Student number: 2948222

e-mail: f.j.kapulla@umail.leidenuniv.nl

Daily Supervisor

Name: Matthijs Tijink

e-mail: Matthijs.Tijink@cqm.nl

Institute: CQM

Function: Senior Data Consultant at CQM

Second supervisor (if applicable)

A second supervisor from the master Statistics and Data Science is needed, when the daily supervisor is no staff member of the departments/Institutes of the master

Name: Dr. Erik van Zwet

e-mail: E.W.van_Zwet@lumc.nl

Institute: LUMC

Function: Associate Professor, medical statistics

EMOS (European Master in Official Statistics) project (yes/no)

Intended independent reader (to be filled in by Statistics and Data Science supervisor)

Each thesis is judged independently by a second member of the Statistics and Data Science organizations. The independent reader is not involved in the research of the project and has no direct hierarchical relation with the supervisor(s).

Name:

e-mail:

Institute:

Function:

1.1 TITLE / ABSTRACT THESIS PROJECT PROPOSAL

Title: Setting up an automatic and robust workflow for building an explorative Random Forest Machine Learning model

Abstract (max 150 words):

Goal of this thesis project is to set up a fairly automatic and robust workflow for building a random forest model which may contribute to the exploratory phases of a project for the data consultancy company CQM in Eindhoven. One major output component of the workflow should be a list of variable effects with corresponding standard errors. For that, simple effect size definitions for main and low-order interaction effects are used and standard errors of those are estimated directly using the jackknife method on the inherent bootstrap samples of the random forest. Other output components are a measure of variable importance or accumulated local effect plots. The former gives information about how important variables are for the model to make predictions whereas the latter method is used to visualize possibly complex, non-linear relationships between predictors and outcome.

1.2 ECTS Justification for the preparation of the Thesis Proposal (max 500 words)

The investment for the study and writing of this thesis proposal should be 4ECTS. Please clarify how many hours were spent on the activities needed for the writing of this thesis proposal:

Write down here your justification of your hours for this thesis proposal, know that you can also include hours spent on further improvement of academic skills (For example presentation workshop / writing workshop or module etc.).

- Thesis lecture on responsible and reproducible research (8hrs)
- Literature research about explainability in the domain of machine learning and artificial intelligence (32hrs)
- Studying concepts of analyzing variable effects of black box models (22hrs)
- Consultations with internal and external supervisors to determine the goals and research question of the thesis and internship. (14hrs)
- Familiarize myself with the working environment and daily procedures during the first week at CQM. (16hrs)
- Setting up a plan for weekly consultation meetings for supervision of the meetings. (6hrs)
- Writing and editing this proposal. (14hrs)

2. DESCRIPTION OF THE PROPOSED RESEARCH PART OF THE THESIS PROJECT

Write a concise proposal of a maximum of 1200 words. This should contain:

2.1 The Research Problem

Introduce the research topic and discuss at least one scientific paper, book chapter or report on the topic

At CQM, a data consultancy company in Eindhoven, a wide variety of projects in the domain of quantitative and statistical analysis as well as data science are covered. They work together with companies such as Philips, ASML, NS or ProRail and often those projects involve observational data such as logged data from complicated machines. A typical goal in those projects is to explain the relation between several predictors and an outcome in order to better understand the underlying domain problem.

This is a typically goal of the data modelling culture in classical statistics where a stochastic data model is assumed for the true functional relationship between predictors and outcome (*Breiman, 2001*). E.g. Linear regression falls under this paradigm which assumes a linear relationship between predictors and outcome variable as well as homoscedasticity and normality. Coefficients are estimated from data and the model then used for information and/or prediction. However, linear regression may often be too simplistic to model complex data structures because of the underlying data and model assumptions it makes.

Contrary, modern machine learning (ML) algorithms such as random forests, support vector machines or neural networks consider the true functional relationship to be complex and unknown. Their approach is to find a function that operates on the input space to predict some response (*Breiman, 2001*). These algorithms are very flexible, complexity controlled by hyperparameters and usually the main goal here is really prediction accuracy. However, this flexibility comes at a cost in terms of less interpretability and transparency (*Apley, 2020*). For those reasons, model-agnostic methods have been developed in the recent years to visualize marginal variable effects on a global level and to explain individual predictions on a local level (*Apley, 2020*).

Nonetheless, it is also desirable to obtain point estimates of variable and low-order interaction effects with corresponding standard errors similar as in linear regression from classical statistics. The rationale for this is to avoid the need of time-intensive assumption checks of the underlying data structure and model assumptions, as it is necessary for e.g. linear regression models, and still get quickly an idea about the size and direction of variable effects. Furthermore, visualizing marginal effect plots of all predictor variables can be visually cumbersome, especially when in practice data sets contain multiple outcome variables, many features and when multiple model variants are used.

2.2 Research aims

The main research aim is to assess the estimation of simple defined effect sizes for main and low-order interaction effects and estimate corresponding standard errors directly when training bagged learners such as random forests. To estimate those standard errors a bias-corrected version of the Jackknife is used that is applied on the inherent Bootstrap samples of the random forest. Wager et al. (2014) implemented such method to estimate standard errors of random forest predictions. Here, this method is manipulated for the purpose of estimating the variance covariance matrix between random forest predictions. By means of doing so it is possible to estimate standard errors of the simple defined variable effect in a random forest setting directly.

2.3 Research plan

Be specific here. You may think of the following parts: Data description (in case real data are used), design of simulation experiment (in case a simulation experiment is performed), methods which you will use to address the research question(s).

For any machine learning model $\hat{f}(\cdot)$, a simple effect size of predictor variable j could be defined and estimated by fixing some point in the middle of the predictor space and shift the coordinate of interest by some amount to form two new points A and B.

- $\hat{A} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_j - k \cdot \hat{\sigma}_{x_j}, \dots, \hat{\mu}_P)$
- $\hat{B} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_j + k \cdot \hat{\sigma}_{x_j}, \dots, \hat{\mu}_P)$

Here, the ML model $\hat{f}(\cdot)$ approximates the true functional relationship between P predictors and an outcome variable and $\hat{\mu}_j, \hat{\sigma}_{x_j}$ are estimates of the true mean and standard deviation of the j -th predictor variable respectively. Hence, \hat{A} and \hat{B} are estimated from data. An effect of x_j on the conditional mean $E[Y|x]$ can be defined as $\frac{f(B) - f(A)}{2 \cdot k \cdot \sigma_{x_j}}$ and estimated by $\frac{\hat{f}(\hat{B}) - \hat{f}(\hat{A})}{2 \cdot k \cdot \hat{\sigma}_{x_j}}$. This variable effect is corrected for possibly non-linear effects in the other variables and local at the mean of all other variables. To obtain a standard error estimate of such variable effect, the following expression needs to be estimated:

$$Var\left[\frac{\hat{f}(\hat{B}) - \hat{f}(\hat{A})}{2 \cdot k \cdot \hat{\sigma}_{x_j}}\right] = \frac{1}{(2 \cdot k \cdot \hat{\sigma}_{x_j})^2} (Var[\hat{f}(\hat{B})] + Var[\hat{f}(\hat{A})] - 2 \cdot Cov(\hat{f}(\hat{A}), \hat{f}(\hat{B})))$$

Here, $\hat{\sigma}_{x_j}$ is considered to be constant and taken out of the variance expression since the variation of this estimate is negligible. A natural estimator for $Var\left[\frac{\hat{f}(\hat{B}) - \hat{f}(\hat{A})}{2 \cdot k \cdot \hat{\sigma}_{x_j}}\right]$ is the following expression:

$$\widehat{Var}\left[\frac{\hat{f}(\hat{B}) - \hat{f}(\hat{A})}{2 \cdot k \cdot \hat{\sigma}_{x_j}}\right] = \frac{1}{(2 \cdot k \cdot \hat{\sigma}_{x_j})^2} \cdot (\widehat{Var}[\hat{f}(\hat{B})] + \widehat{Var}[\hat{f}(\hat{A})] - 2 \cdot \widehat{Cov}(\hat{f}(\hat{A}), \hat{f}(\hat{B})))$$

In general, estimating the standard error of random forest predictions ($\sqrt{\text{Var}[\hat{f}(x)]}$) can be challenging, since there are two distinct sources of noise. Firstly, sampling noise and secondly Monte Carlo noise arising from the use of a finite number of bootstrap samples (Wager, 2014). However, Wager et al. (2014) implemented two methods to estimate standard errors of random forest predictions. A bias-corrected version of the Jackknife-after-Bootstrap and the Infinitesimal Jackknife. For this project the former approach is used which applies the Jackknife on top of the inherent Bootstrap samples that were used to fit the random forest. In order to estimate the variance-covariance matrix between random forest predictions, this method is slightly manipulated.

In general, the Jackknife-after-Bootstrap estimate (Efron, 1992) is defined as follows:

$$\widehat{\text{Var}}_J^\infty = \frac{n-1}{n} \sum_{i=1}^n (\bar{t}_{(-i)}^*(x) - \bar{t}^*(x))^2$$

Here, J stands for the Jackknife-after-Bootstrap method to estimate the variance of a random forest prediction, assuming an infinite size of bootstrap samples (∞). $\bar{t}_{(-i)}^*(x)$ is the expected random forest prediction for x, using only those trees where the i-th observation was not part of a bootstrap sample (out-of-bag), while $\bar{t}^*(x)$ is just the average over all tree predictions or in other words the expected random forest prediction for x. However, in practice one can only work with a finite number of bootstrap samples. The natural Monte Carlo approximation to the estimator above is:

$$\widehat{\text{Var}}_J^B = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(-i)}^B(x) - \hat{\theta}^B(x))^2, \text{ where } \hat{\theta}_{(-i)}^B(x) = \frac{\sum_{\{b: N_{bi}^*=0\}} t_b^*(x)}{|\{N_{bi}^* = 0\}|}$$

N_{bi}^* indicates the number of times the i-th observation appears in the b-th bootstrap sample. Hence, for each observation i it is checked in which bootstrap samples that observation was not part of. Based on the corresponding fitted trees, predictions are made for x and averaged to obtain $\hat{\theta}_{(-i)}^B(x)$. This value is then subtracted by the random forest prediction $\hat{\theta}^B(x)$ and the difference squared. To estimate the covariance between two random forest predictions for points x and x', $\widehat{\text{Cov}}_J^B$ can be manipulated as follows:

$$\widehat{\text{Cov}}_J^B = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(-i)}^B(x) - \hat{\theta}^B(x)) * (\hat{\theta}_{(-i)}^B(x') - \hat{\theta}^B(x')),$$

$$\text{where } \hat{\theta}_{(-i)}^B(x) = \frac{\sum_{\{b: N_{bi}^*=0\}} t_b^*(x)}{|\{N_{bi}^* = 0\}|}$$

In order to assess the reliability and variability of the simple defined variable effect and corresponding standard error estimates the following simulation set up will be implemented:

- Two different scenarios of true functional relationships between predictors and outcome are considered. One only includes linear and the other non-linear terms.
- Observations belong to different groups and each observation has one single outcome.

We use a balanced group design meaning that all groups are equally large. However, the number and size of groups across scenarios are varied.

- Variables are multivariate normally distributed with $\mathcal{N}(0, \Sigma)$
- Each group has a random intercept which is i.i.d. normally distributed with $\mathcal{N}(0, \sigma_u)$
- Residuals are i.i.d. normally distributed with $\mathcal{N}(0, \sigma_\epsilon)$

Another output element the workflow may contain is a permutation variable importance measure. The rationale of this is that permuting a predictor breaks the association between the feature of interest and outcome (Altmann, 2010). It is important to mention that such measure does not reflect the intrinsic predictive value of a feature by itself but how important this feature is for a particular model. Hence, to obtain meaningful insights from such measure it is crucial to train a model that generalizes well on unseen data (Altmann, 2010).

In order to visualize possibly complex, non-linear relationships between predictors and the predicted outcome variable, accumulated local effect (ALE) plots can be used. ALE is a type of global model-agnostic method which shows the marginal effect of a predictor on the predicted outcome of any ML model (Molnar, 2020). ALE plots are a different variant of well-known partial dependence plots (PDP). However, one disadvantage of PDPs is the assumption of independence. It is assumed that variables in the set of predictors for which the partial dependence is to be computed do not correlate with variables of the complementary set. This is because predictions are averaged over the marginal distribution of all the other variables (Greenwell, 2017).

2.4 Expected Results/end product:

List the expected results (e.g. software package, expected results of a data analysis, a concrete advice to a specific stakeholder, results of simulation study, results of methodological research).

Programming language R and packages ‘ranger’ and ‘pdp’ will be used to implement the methodological ideas described above. The end product will be a simulation study where the estimation of variable effects and corresponding standard errors is evaluated for different scenarios (linear, non-linear, interaction effects etc.). Here, it is expected that on average the direct standard error estimates of the simple defined variable effect should be close to the standard deviation of the variable effect estimates. Furthermore, a workflow will be built where the researched methods of the simulation study come into practice as well as a few other metrics (permutation variable importance and ALE plots) that add to the explainability and interpretability of a fitted random forest model.

2.5 Reference list:

Cite and list the cited literature into a format according to a common citation style guide of the field of research (APA, MLA, Chicago, IEEE, ...)

References

- Altmann, A. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics* 26.10, 1340-1347.
- Apley, D. W. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B*, 1059-1086.
- Cafri, G. a. (2016). Understanding variable effects from black box prediction: Quantifying effects in tree ensembles using partial dependence. *Journal of Data Science*, 67-95.
- Greenwell, B. M. (2017). pdp: An R package for constructing partial dependence plots. *R J.* 9.1, 421.
- Molnar, C. (2020). *Interpretable Machine Learning*. Lulu.com.
- Trevor Hastie and Zhao, Q. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics* 39.1, 272-281.
- Wager, S. T. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* , 1625-1651.

3. ACTIVITIES TO GET EXPERIENCE AS A WORKING STATISTICIAN/DATA SCIENTIST IN A POSSIBLE FUTURE WORKING ENVIRONMENT

Formulate, together with your supervisor(s), three concrete working activities in which you will participate during your thesis project. One of these activities should involve developing communication skills. A list of possible activities can be found on Brightspace and in the information for supervisors, but it is allowed to formulate other activities. Give for each of the activity a short description and a short motivation for choosing the activity.

Activity 1 Actively participate in meetings at the working place

Every Monday morning I will participate at weekly meetings of my department at CQM. During those sessions every employee gives a quick overview about current projects and personal working goals of the coming week. Next to work-related topics this is also a nice opportunity to share interesting events/experiences that happened throughout the prior week. This can be totally unrelated to work and helps to have a joyful start for the week.

From those meetings I expect to get insights about the kind of projects my co-workers are dealing with. Furthermore, I want to set myself clear weekly goals through those meetings. In that way I hope being able to achieve my overall thesis goal by taking small, continuous steps.

Activity 2: Causality & AI for Health meeting (November 22)

At this meeting researchers working on methodology of causal inference are brought together in the context of the AI Convergence between Leiden, Delft and Rotterdam (<https://convergence.nl/ai-data-digitalisation/>). The goal with those meetings is to get to know each other's work in this area and to foster relationships between researchers working on causality within different fields and across the three universities. Personally, I hope to get insights about the most recent developments in that field since it touches upon my thesis topic. Furthermore, I hope to get in touch with researchers and possibly obtain more ideas for my project in terms of what I can include into the workflow for building an explorative machine learning model.

Activity 3 (developing communication skills): Interview a person other than your direct supervisor on the opportunities and limitations of the working place

In the course of my thesis internship I will interview a person other than my direct supervisor on the opportunities and limitations when working for CQM. Personally, I plan to go into industry after my study. However, I am very much open to work in the domain of consultancy, data science or any other field related to data analysis. Hence, it will be very useful to get direct insights whether data consultancy would suit me well in the long run.

4. WORK PLAN AND SUPERVISION

4.1 Supervision:

Describe the arrangements regarding the type and frequency of meetings between student and daily supervisor(s) and on roles and responsibilities. If there is a second supervisor from Statistics & Data Science, also describe type and frequency of communications between the second supervisor, the daily supervisor and the student.

Every week Tuesday I will meet with my external supervisor from CQM to update each other about the progress of the thesis. Apart from that, it is always possible to ask my external supervisor questions in person on short notice since we work most of the time at the office. Furthermore, I will also meet at least once a month with my university supervisor Erik van Zwet to check the thesis progress from an overall scientific point of view.

4.2 Time Schedule

Carrying out the thesis project should take 30 EC (exclusive 4 EC to write this proposal). Present a feasible time schedule of your activities. Note that 30 EC corresponds to 21 weeks full time work. Make a detailed plan (week by week) so that at each supervisor-student meeting, it can be discussed if things are still going as planned, and if not, how to tackle that. Be aware that writing takes time. Indicate what elements can be cut / reduced if necessary.

Week	Activity	Week	Activity
1	Introduction week at CQM	14	Creating workflow for building ML model and Midterm progress analysis with supervisors
2	Defining Methodology of Thesis	15	Create workflow for building ML model based on simulation results
3	Writing Proposal	16	Write down first draft of workflow implementation
4	Writing down first draft of Methodology Part	17	Apply workflow to real data set
5	Defining simulation setup and writing first draft of simulation description	18	Time for other internship-related tasks
6	Setting up Simulation in R	19	Revise workflow draft and write

			<i>down workflow results for dataset</i>
7	<i>Setting up Simulation in R</i>	20	<i>Room for workflow additions</i>
8	<i>Get Simulation in R running</i>	21	<i>Room for workflow additions</i>
9	<i>Christmas and New Years vacation</i>	22	<i>First draft of Introduction</i>
10	<i>Get Simulation in R running</i>	23	<i>Improving text. Discussion draft</i>
11	<i>Analysis and Visualization of Simulation Results</i>	24	<i>Improving text, prepare internship presentation and administrative tasks</i>
12	<i>Revise Simulation draft and add simulation results</i>	25	<i>Improving text, prepare internship presentation and administrative tasks</i>
13	<i>First draft of data description</i>	26	<i>Final editing, review and presentation</i>

4.3 Infrastructure

Describe the arrangements offered to the student to facilitate the students' work progress (For example, guest employment, a desk, shared office, computer, access to a computing server)

At CQM, interns do not have a fixed working desk which has the advantage of working in company with different employees. Furthermore, I got my own CQM laptop where I have access to the network paths of the company. As agreed upon, I come four times per week to the office and one day I will work from home.

4.4. Other Courses / Activities:

- What courses (how many ECTS) still need to be obtained during the thesis project before graduation. Please adapt your time schedule to incorporate this.

I take the course Text Mining next to my thesis internship. However, I do this on a voluntary basis since I already obtained all ECTS required in terms of courses.

- Are there other reasons that may make it impossible to spend ± 21 consecutive weeks on the thesis project
No, there are no other reasons.




5. AGREEMENT PAGE

Student and Supervisors hereby declare that they agree to the arrangements in this proposal.

Student declares that he has provided the supervisor with the Documentation for Supervisors

The student hereby declares that both this proposal, and its resulting thesis, will be free of plagiarism (cf. Rules & Regulations of the Board of Examiners).

The supervisors and student hereby declare that they have applied, and will apply good scientific practices, that follow the [University Academic Integrity Regulations](#)¹ and the [Ethical Guidelines from Statistical Practice](#)². When in conflict with each other, the [University Academic Integrity Regulations](#) should be followed.

	Name	Signature
Student:	<div>Date 24.01.2023 Felix Kapulla</div>	 
Supervisor:	24 january, 2023	
Second Supervisor:	<div>27-01-2023 Matthijs Tijink</div>	

This signed proposal should be submitted to the Thesis Committee (thesis@stat.leidenuniv.nl)..

Thesis committee member:

¹ <https://www.universiteitleiden.nl/en/research/about-our-research/quality-and-integrity/academic-integrity>

² <https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx>