UNIQA Data Science Challenge 2022

Start: 1st March 2022
End: 31st March 2022

Announcement of the winning projects (during a virtual event):
In the second half of April 2022 (exact date to be announced at the beginning of April)

In this first UNIQA Data Science Challenge of 2022, we want to explore pricing!

When building their insurance tariffs actuaries around the world need to create many complex models using GLM/GAMM techniques. Typically, when predicting insurance risk, they try to model the expected number of claims ("loss frequency") and the average claim amount ("loss severity") separately.

Such models need to be both accurate and transparent.

Most complex models can have hundreds of transformations of features, so there is always the question: have we found all relevant ones? Or how much more is still unexplained by the model relative to the best possible model?

We provide you with the dataset and the output of our GLM model (as one of the variables). Your task is to propose a XAI-class (explainable AI) model that performs better than our GLM model. Your model should be both accurate and needs to be able to explain its predictions (per observation).

The dataset provided consists of the coded variables used in our valuation models. To simplify the task, we reformulate the problem to classification, i.e. the probability of claim is modelled instead of the frequency of claims.
To be more specific, what is usually predicted is the frequency of the claims. Since a client can have more than one claim, a frequency of 0.5 does not necessarily mean a probability of 50% of the client having a claim, but it means on average the client will have 0.5 claims per year.
In this challenge we reformulated the problem to a pure classification problem meaning you fit a 0/1 variable, where 1 encodes that the client had at least one claim. Your model will predict a probability for each client having at least one claim.
(Note: The explanatory variables of the client also contain our GLM fit. This fit was created using more fields than provided to you. This was done with the idea to encourage you to include the GLM fit in your model in some way. However, it is not mandatory to include it.)

So, your goal is to use a model of your choice to predict a probability of having at least one claim for each customer given their characteristics! The quality of the fit will be assessed using the Gini coefficient which contributes to 1/2 of the points. The remaining 1/2 of points will be given for simplicity, explainability and reproducibility, see section evaluation.

Show us your programming skills in Python or R by doing your entire analysis in a Jupyter notebook and your ability to present your findings in a clear and structured way by explaining your analysis

and results in a summary of maximum 3 pages. An explanation of the variables used should be provided both per observation and at the aggregate level.

Participation

The competition is open to all interested persons of legal age which can participate individually or as a group. Groups may consist of a maximum of 3 people. All participants competing for the main prizes must be enrolled at a European university with a valid university email address or proof of current study at the time of submission and must not exceed the age of 30. All other participants who are either over 30 and/or do not have a valid mail address (or proof of current study) of a European university are welcome to take part in the challenge but out of competition for the cash prizes.

Requirements for the submissions

- The dataset to be used for the analysis will be made available on Dropbox. The link will be published **on our web page at the beginning of the challenge (1st March 2022).**
- The analysis must be performed and submitted in R or Python using a Jupyter notebook.
- The Jupyter notebook must run as is on any computer.
- Every part of the analysis must be reproducible. The packages used for the analysis should be listed as a part of the code header, commented out.
- The 3-page summary must be submitted in pdf format. This summary should contain a description of the solutions obtained, and an explanation of the approached and the results achieved.
- The upload of the summary and the notebook must be done together with the specification of an email address on https://www.uniqa4ward.com/en/challenge.html
- All submitted documents must be written in English.
- Participants must use the provided dataset, using other datasets is not allowed.
- All persons submitting a joint analysis may not submit individual solutions.
- The submitted analysis must contain the name, the nationality, the phone number and the university mail address of all involved persons.
- A main contact person must be specified in case of a joint submission.

Evaluation

The analyses will be evaluated blindly (i.e., without knowledge of authorship) by a jury based on two different criteria which are weighted equally:

- **1/2:** Model predictability measured by Gini coefficient.
- **1/2:** Here we focus on the more conceptual qualities of the model, more specifically:
    - **Explainability:** (per observation, and also, you should be able to draw conclusions about possible extensions/improvements to our GLMs/GAMMs).
    - **Simplicity:** (simpler solution is better, easier to adjust to other models in the future, easier to use and maintain). Also, fewer models are beneficial, i.e. avoidance of model stacking
    - **Reproducibility**: can the model easily be generalized into a “tool” which can be used for other datasets?

For the first part, the model predictability, the Gini coefficient will be followed objectively, and the jury will merely play an administrative role. For the second criterium, it will actively evaluate the submissions.

The jury consists of at least 5 members, nominated by the Pricing International department of UNIQA. One member of the jury will be appointed as the chairman of the jury. In the event of a tie, the chairman's scores will be used to determine the ranking. From all submissions that meet the

requirements to compete for the main prizes, the jury will select the three projects that have the highest score according to the above criteria and invite the winners to a final (online) event. At this event the first 3 places will be announced, and the winners will have the opportunity to present their analyses.

Prizes

The following prizes will be awarded to the first 3 places among all participants who meet all the requirements as specified above:

1st place 1.500 EUR
2nd place 1.000 EUR
3rd place 500 EUR

Additionally, the participant with the “most predictive model” (first criterion of the evaluation) will be awarded a trip for two persons to the 2022 **WeAreDevelopers World Congress** in Berlin (https://www.wearedevelopers.com/world-congress#/), including:

- 2 tickets for the Congress on 14-15 June 2022
- Economy flight ticket or equivalent train ticket depending on the location of the winner
- 2 nights’ stay in a 4 stars hotel (double room) with breakfast included

Obligations of the participants, disclaimer, and additional rules

- By participating, participants warrant that they are the author of the analysis and do not infringe the intellectual property rights of any other person or entity. In particular, authors must own the rights to all illustrations, images, and elements that they use in their submission.
- By submitting their analysis, entrants consent to the use of their name in the announcement of the award winners without further restriction.
- Claims for compensations of the participants are excluded.
- If for any reason the competition cannot be carried out as planned, the organizer reserves the right to cancel the competition at its own discretion.
- Submissions that do not comply with all rules will be disqualified at the discretion of the jury.
- The term organizers refer to UNIQA
- By uploading a submission, participants agree that their provided email address may be used by the organizers for announcements of further challenges.
- By participating, participants grant the organizers free of charge a perpetual, exclusive, non-transferable license to publish, display, reproduce, use, share, edit, reproduce and create derivative works of their analysis in any media, in whole or in part, for any purpose.
- By participating in the Challenge, participants agree to provide, upon request, documents that allow the organizers to verify the age and authenticity of the university email address.