# Bayesian Statistics and Probabilistic Programming

## *Spring 2021*

# Assignment 02

See the Learning from Imbalanced Insurance Data (Cross-sell Prediction) dataset from Kaggle. This is a binary classification problem to apply **Logistic** or **Probit regression**, possibly with some quadratic or interaction term, with techniques/ languages in the course: JAGS, Stan, Greta, MCMCpack, `brms`, `rstanarm` (do not use them all, but more than one). You should evaluate the relative importance of predictors by resorting to any variable selection method or shrinkage priors you deem appropriate. Do try to document everything.

Like with most datasets, Kaggle stores several takes on these data, either from the exploratory (visualization) or predictive point of view. You can, and should, read them, borrowing suitable ideas –with due reference. Some of the "frequentist" predictive methods there *(e.g. Random Forests)* have Bayesian translations, absent in our course on account of time constraints. Obviously it is not demanded (nor banned) that you use them. If you do, they may provide a touchstone, either in their Bayesian or the Frequentist avatar. Results in the Kaggle website seem to suggest that the best classification is obtained with the *Ensemble methods* family (varieties of Boosting or Random Forests). Nonetheless, Logistic regression appears to give quite decent predictions.

This is an open assignment, you can go beyond the bare minimum stated above. This is especially true concerning the ***imbalanced*** qualification in the problem title: this is a common characteristic of classification problems with credit approval or insurance claims data: typically only a small fraction of observations belong to the *credit defaulted* or *insurance claimed due to sinister* class. Such training sets tend to distort predictions and, given their relevance in Insurance or Banking applications, a number of correcting devices have been designed to address this particular difficulty. Again, I do not expect you use them in your submission (but keep in mind the problem exists, should you find it in your professional life). On the other hand, these methods (e.g., SMOTE for unbalanced classification) can be the subject of a course project, if you feel so inclined.