Cladag2019 data science competition

powered By TIM

On the occasion of CLassification And Data Analysis Group (CLADAG) 2019 Conference, hosted at the University of Cassino and Southern Lazio on September 11-13, we announce the “CLADAG2019 Data Science (DS) Competition”, in collaboration with TIM. In particular, TIM will support the organisation of the competition and it will provide the data set.
The aim of the data science challenge is to promote the application of the most recent methodologies proposed in the scientific literature to tackle the data-analytics challenges that real-world operating companies have to face throughout their activities.
The challenge is for individual researchers or for teams of up to three researchers that will be asked to provide a solution to a real data analytics/machine learning problem, that is, prediction and analysis of customer churn, with focus on fixed telephony customers. The participants should have a compatible skill-set and background; no restrictions on the participants job position.

CLADAG2019 DS competition will assign 2 prizes:

1) 2000 euro for the overall best performing model in terms of prediction.
2) 1000 euro for the best description of the data, and for the best interpretation of the obtained results, that may provide new insights on the phenomenon.

The grants will be paid as funds to the institutions that winners belong to, or, if the winners prefer, they will be paid directly, in this case all taxes on the prize are the sole responsibility of the winners.

Competition general info

The customer churn prediction is key for TLC companies: to the loss in revenues due to customers leaving the service, is added an increase in marketing costs to acquire new customers.
Therefore preventing/reducing customer churn is important, but the competitive advantage comes from the ability of the company to identify: i) subsets of customers that are more likely to leave the service, ii) the causes of customers churn. In fact, the goal is reducing the customer churn while not lowering the service prices, and this can be done by specific customer care policies, tailored on customers that are most likely to churn.

Goal of the competition

The data scientists entering the competition are required to build a classification model to predict the risk to churn of fixed telephony customers and to identify the discriminant variables that best describe the customers with higher churn risk. Participants will build their churn prediction model using the data provided by TIM, but they can integrate the available feature set with further information gathered from open data sources.

Competition specifics

Each team registering to the competition will nominate a team leader; it is not possible to change the team's composition once registration for the competition has been completed. The team leader will receive the link to access the data set to analyse. In particular, the data will be provided by TIM in a pseudo-anonymised form and will refer to past TIM customers: a fraction of the data will be released by TIM unlabelled and it will be used as a validation data set to assess the performance of the competing models. Participants will have access to the data once the competition starts: then they will have three weeks of time to develop their proposal.

Each team will have to provide:

- a vector of predictions referred to an unlabelled test data set;

- a report of up to ten pages should be provided to describe the analysis strategy, the methodology used and the interpretation of the results, highlighting the main findings.

- a script in Python or R programming meta-languages to guarantee the reproducibility of the obtained results;

As mentioned above, the set of features/predictors can be integrated by using open data libraries from official statistics provided by ISTAT or from the Italian Government website, or any data source publicly accessible. Extant further data sources can be used either to enhance the predictive performance of the model or to increase the interpretability of the results.

Project evaluation

The scientific committee will consist of members from SIS (Italian Statistical Society) and from TIM data scientists, and it will decide upon the best teams. In order to be evaluated, however, the proposed projects will have to meet minimum performance requirements in terms of prediction accuracy. The projects evaluation phase will last three weeks. Eventually, a shortlist of top performing teams will be invited to present their projects in the final stage of the competition, that will determine the price winners.
The committee evaluation will be focused on the following aspects:

- The classifier prediction accuracy, measured by ROC-AUC on the validation set retained by TIM.

- Quality of the report in terms of both formal rigour and readability.

- The process of model selection and tuning.

- Interpretation of the results.

- The contribution of external features to the prediction and/or to the interpretation of the results.

Important dates

- **13/9/2019** data science competition kick-off presentation at Cladag2019 and official start;

- **2/10/2019** deadline for participants: project submission;

- **23/10/2019** shortlist of top teams announced;

- **tba** Final presentation and winners announcements.

How to register

To sign up for the competition the team leader is required to fill the registration form at the following link upon acceptance of the legal conditions that apply.

The link will be active since **13/09/2019** and it will expire at **23:59** of **1/10/2019**.

A registration form that contains false information about one of the members of a team, or members registered into more than one team, will cause the exclusion from the competition of the involved team(s).