Course: Machine Learning with Python

Authors: Daniel Borek, Maria Cristina Hinojosa Lee, and Maarten Velghe

# Executive summary

This work corresponds to the final project of the course Machine Learning with Python offered by the Institute for Continuing Education in Science. For it, different techniques of Machine Learning were applied using a dataset with information over the guests of a Hotel, to predict the behavior of potential customers. The activity was divided in five points: preparing the data set, predict the projected revenue per clients, predict which clients will cause damage and the damage that they would cause, and create a measure of the expected value for each potential customer and select two hundred of them.

For the data processing basic visualization techniques were used to get acquainted with the data. Missing data was dealt with mode and mean imputations. Then the original features and one dummy were dropped and the data was rescaled.

For the second question related to the projected revenue per client, the Gradient Booster algorithm was selected. To try to white boxing the algorithm, variable importance based on impurity reduction and permutation importance were used. The strongly predictive features identified were: Profit per night, Mean taxable income of neighborhood, Presidential suit previous use, and the amount spent in shop.

We tried GradientBoosting Clasiiffier and HistogramGradient boosting. The fancy model are good, but choosing the subset of the variables improved overal performance that the simple logistic regression on subset of data (we choose features basing on Perason correlation and mutual entropy). We tested a bunch of basic models, the logistic was the best. Probably if we would get better results if used the algorithms mentioned in the first part on the data with removed features.

For predicting abount we tried XDBoost regressor we decided to stick to Gradient Boosting for regression. We were able to improve initial metrics by selecting subset of variables,

For final model we used binary outcome [0, 1] not proababilites and it may negatively influence our prediction, But much more important is the prediction of the money the client will left.

Overal we improved all metrics from the example notebook. Maybe better solution would to predict overall amount of profit for the hotel from the start (as classification of good or bad client, doing it the way we did we loose some information about uncertainty)S

# Deliverables

At the end, you will present 4 documents:

1. an executive summary of *no more than 1 page*, preferably in Word or PDF.
2. a technical report, containing information on the most important steps in your analysis: what did you do, why, and what were the results. This document mixes plain text, relevant output from your code and essential code snippets. This can be in Word, PDF, or notebook, plain documents should *not be longer than 5 pages*.
3. your full code, as a notebook containing code, comments, and output.
4. the final list of selected clients, with their predicted revenue, predicted damage status (yes/no), predicted damage costs, and overall predicted revenue, as csv.

When delivering notebooks, please provide them in **both .ipynb and .html format**. You can also combine 2) and 3) (the full code and technical report), but make sure to include sufficient comments, with information on your strategic choices as a modeler and your reasoning. Note also that at all times it should be clear what belongs to your final analysis, and what pieces are merely try-outs that are not part of the final approach.