HA Data Science Internship

Assessment

October 2019

The purpose of this assessment is to evaluate the ability of the candidate to train and test a basic machine learning model in the fraud detection domain. The candidate is provided with a dataset (in csv format) and with a brief explanation of the problem and the contained features. The final goal is to provide a prediction for some unseen samples using the trained model. The unseen samples might have been available (and therefore collected) in a different moment from the ones in the training set, and some features might be missing, being related to the labeling process itself.

Even if the structure of this assessment is very similar to the one proposed in data science competitions (eg. Kaggle), the final criterion used to evaluate the solutions isn't a ranking of all the received predictions, but the clarity, scalability and completeness of the provided solution. Obviously, *whichever* metric* the candidate employs to score its model: the higher, the better; but again: do not lose your mind in achieving a 0.001 improvement in accuracy selecting that specific hyperparameter for the latest state-of-the-art model found on Github and still in *beta*, but rather focus on providing a solid explanation of the rationale behind the (possibly) engineered features and the metric(s) employed for scoring the model.

Feel free to use any library you want as long as it is freely available on the web, or developed by yourself. Python is strongly preferred.

* the metric that the candidate will decide to maximise for the specific task and, above all, the rationals used for the decision are part of the assessment and differ from what happens in Kaggle-like competitions as it is not provided in the problem formulation. Pick a metric, describe why you picked it, and declare the model performances using it. Deliver the predictions.

Dataset:

The training set (HA_Data_Science_train.csv) is composed by 16762 samples, each sample contains information about a listing on our platform. Goal of the classification is distinguishing between scam and legit listings at creation time, where the majority of these features are actually available. A brief explanation of the available features is provided in the subsequent table.

Feature Name	Description
LISTING_KIND	0 - entire place; 1 - private room; 2 - shared room
LISTING_CITY	The city where the listing is located
LISTING_PRICE	The monthly rent (€) of the listing
IS_ARCHIVED	If the advertiser creating the listing has been

	archived or not.
ARCHIVE_REASON	The reason why the advertiser creating the listing has been (possibly) archived.
LOGIN_COUNTRY_CODE	The (last) country where the advertiser logged in from.
LISTING_COUNTRY_CODE	The country where the listing is located.
LISTING_REGISTRATION_POSSIBLE	If it's possible to use listing's address for registering at the city's municipality.
ADVERTISER_COMPLETENESS_SCORE	Percentage of completeness of the advertiser's profile.
MANAGED_ACCOUNT	If the advertiser creating the listing is managed by our employees or not.
HAS_PROFILE_PIC	If the advertiser creating the listing has a profile pic or not.
BROWSER	The browser used to create the listing.
OS	The operating system used to create the listing.
ANONYMISED_EMAIL	The email address (anonymised) of the advertiser. Letters have been changed with random letters, numbers have been changed with random numbers, all the other characters have been maintained. The email domain has been maintained.
IS_SCAMMER	Whether the listing is a scam (1 - bad) or not (0).

The samples for which the candidate has to provide predictions for the target value **IS_SCAMMER** can be found here.

What to deliver:

- A csv file named `HA_Data_Science_test_labeled.csv` containing all the samples in `HA_Data_Science_test.csv`. The values for the column `IS_SCAMMER` (0 or 1) have to be populated using the trained model.
- The Jupyter Notebook (.ipynb) or the Python script (.py) developed for populating the dataset of the previous step.

- *Optional:* A document containing a description of the work, an explanation for the (possibly) engineered features and for the rationales behind the choice of the specific metric. A well written Notebook or a well commented script are also fine.