Exam Machine Learning – IPVW-ICES 2021

vanrompayebart@gmail.com

Due date: Sunday July 4th, 2021

Context – Drunk Russians

Among all international hotel guests, Russians are burdened with the upkeep of a singular reputation: they are (supposedly) the rowdiest bunch one can entertain, and are equally well-known for unbridled spending as for racking up extensive costs in damages to hotel infrastructure, staff, and occasionally also other guests – costs which typically cannot be recovered once the guest has sought out the safety of his (or her) homeland.

It is your job as a data scientist to screen applying Russians clients for an exclusive hotel in the Bahamas - yes, it's the kind of hotel you need to apply for!

The data

At your disposal is a training set containing data about the behavior of 5000 Russian hotel guests (train_V2.csv). This data set contains information about the profit the hotel made during their last visit (excluding damages), but also whether they caused damages during their last visit, and for what amount. These outcomes are respectively called 'outcome_profit', 'outcome_damage_inc', and 'outcome_damage_amount'. To predict them, you have access to a host of personal information: previous history of profits and damages, use of hotel facilities, socio-demographics and behavioral scores from the staff of other hotels within the hotel chains. A minor description of features is available in dictionary.csv.

You also get information on the 500 applicants for the 2021 season (score.csv). It is your job to return a list of 200 clients that offer an attractive balance between projected profit for the hotel, and anticipated damages.

You will notice the data set contains a large number of oddities and unclarities. You are expected to provide some minor reflection on these in your technical report, and to clearly draw conclusions and how these translate into acceptable technical approaches that you will apply.

Possible approach

To generate a client list, you can (but don't have to) follow the next steps:

1) prepare the data set
    - briefly survey the data
    - deal with data issues:
        - appropriately handle categorical data
        - treat missing data
        - identify outliers, and choose whether or not to make your analysis more robust by removing these
2) predict the projected revenue per clients

- choose an algorithm, and train it in an optimal way
- score the 500 applicants

3) predict which clients will cause damage
    - choose an algorithm, and train it in an optimal way
    - score the 500 applicants
4) for those that will wreak havoc, predict the amount of damage they will cause
    - choose an algorithm, and train it in an optimal way
    - score the 500 applicants
5) create a measure of the expected value of each applicant, and create an optimal selection of 200 guests

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.

Group work

You are strongly encouraged to work in groups of at most 4 participants. This exam is intended to be a learning experience, and the ability to cooperate, discuss technical topics, and organize analytical work across multiple people is a vital skill for any data scientist. I really believe that talking about data science is grossly underestimated as the best learning approach – another reason why I dislike digital classes ☹ Furthermore, grouping yourself will also make it feasible for me to give you a rich feedback.

Timeline

The data will be posted on Ufora. Your solution should be delivered at the least before **Sunday July 4th, 2021, 23.59** to vanrompayebart@gmail.com (earlier submissions receive unfairly favorable treatment).

You will receive feedback in the form of a grade, and comments to your solution, at an unspecified moment (which without any doubt will be asap).

Note: a little while after posting the exam, an example notebook will be posted. This will contain a very brief example of how one could approach a technical solution. The main intention for this is to

remove all ambiguity concerning outcomes etc, and to show how to deal with at least a few essential data issue in Python.