Long Assignment 2021/2022

Alipio Jorge, Inês Dutra

November 2021 (version 21.11.21)

Objectives

The main goal of this work is to obtain insights and good prediction models for an online buying scenario.

Dataset

The data to be used has been collected by a company during its operation. The company sells luxury items through its website and wants to improve success in its sales and better understand the behavior of the customers/users. For each user interaction, the company records values for a number of attributes possibly related to the success of the sale in a given e-commerce sesssion. When the client buys the product, the variable `bought` has value 1, otherwise it is 0. A number of other variables describe the session.

Below you find a brief description of the variables (data dictionary).

- session_id -> Identifier of the session
- customer type -> Identifies if a user has purchased before (customer/prospect)
- device group -> Device Group (app/mobile web/desktop/etc.)
- visitor type -> If the device is new in the company platform
- has_listing -> Flag that indicates if the session has a listing view
- has_used_search -> Flag that indicates if the session has a search view
- has_recommendation -> Flag that indicates if the session has a recommendation view
- has_add_to_wishlist -> Flag that indicates if the session has an add_to_wishlist view
- has_add_to_bag -> Flag that indicates if the session has an add_to_bag view
- duration -> Session duration (in seconds)
- view_qty -> Number of views in the session
- unique_product_qty -> Number of distinct product page views within the session
- unique_browse_designer_qty -> Number of distinct designer_id within the session
- unique_browse_category_qty -> Number of distinct category_id within the session
- is_subscribed -> If the user is subscribed in the newsletter
- browser_name -> User agent browser name
- country -> Session client country
- bought -> Flag that indicates if the session has an order

You are given the following datasets: - a train dataset with the original proportion of the cases. - a balanced version of the training dataset. - a challenge sample without the target that you will use to participate in a challenge (maximising predictive ability on unknown data)

Guidelines

This data science problem should be approached by following the CRISP-DM methodology (http://jbusse.de/2019_ws_dsci/crisp-dm_phases-tasks-outputs.html). You have to understand the business problem, propose success criteria and see how it can be translated into a machine learning problem. Then you look at the characteristics of the data and you perform the required explorations, visualizations and transformations. Next step is to identify insights, develop predictive models and to evaluate them in order to validate if they

are helpful in the business problems. During the whole process take notes, always identify the questions you want to answer and think before you act: "why is this plot or this transformation useful". You can perform some operations just for the sake of training but you should be aware of that.

The result is a **report** in the form of a **notebook** with clear explanatory text and code that works showing results. The report should be clear, as concise as possible and it should be easy to read and to follow. You will be telling the story of your approach to this problem, so it should have a good narrative flow. Always explain what you are doing, why you are doing it, what are the results and what do you take from those results.

Suggested structure

A report containing:

1. Business understanding
   - Give your view of the business problem following the CRISP-DM list of outpus when adequate.
2. Data Understanding
   - Looking at the raw data, describe variables according to their types: interval-scaled, binary, nominal, ordinal, ratio-scaled. Be aware that there are specific methods suitable to each type of variable.
   - Perform a preliminary analysis (summaries, spread measures, histograms, boxplots, density). These are interesting to be applied to the raw data to "uncover" inconsistencies, outliers, duplicates etc.
   - Perform bivariate analysis (correlations, regression)
   - Provide any insights about the data and the problem that you may have found.
3. Data Preparation
   - List of main changes that can need to be performed to the raw data, including feature selection.
   - Describe the potentially useful ones and their results in terms of data.
4. Modeling: consider the balanced and the non-balanced versions of the dataset as 2 separate problems. Fisrt work with the balanced data and then with the non-balanced data. Try each of the methods below, select hyper parameters using default values and empirical analysis. Separate a test set and use cross-validation on the rest of the examples. Visualize models when possible, visualize results, produce aggregating tables with good insightful summaries of the results, and whatever other tools you may find useful.
   - Nearest neighbor
   - Bayesian Classifier
   - Decision Trees
   - Tree ensembles
   - Support Vector Machines
   - Neural Network Classifier
   - Comparison
5. Evaluation and Main Conclusions
   - What is the best model and the recommended data science procedure for the business?
   - What do you think that the business can gain from your data science effort?
   - What are the lessons learnt?
   - What is your summary of the achieved results?

To submit:

- a fully operational Rmd document or a Jupyter notebook with the selected experiments as clear and concise as possible. Avoid output dumps. Recall that the report is going to be evaluated by your very busy professors and that they may have to skip many pages if your report is too long. Always highlight your best results. Please note:

  - The objectives for each experiment and plot should be clear so that the reader understands why it is worth to read a particular part.
  - The conclusion should be a short high level account of what was observed.
  - It is **not necessary to describe the methods** (unless requested, but you should know their concepts and how they work). It is more important to point out the differences in the methods and the reasons for the results in terms of methods characteristics.
- A 5 minute video (or link to a video), per element of the group with a recorded presentation of the respective part of the work. The presentations of the group, when combined, describe the whole of the group's work.
- The project slide presentation.

Evaluation

- This assignment is worth the values described in sigarra, according to the course you are following.
- Components
  - Report 30%
    - Narrative 10%
    - Writing style 10%
    - Presentation 10%
  - Technical 70%
    - Diversity of the results for the experiments 20%
    - Correctness 30%
    - Challenge performance 10%
    - Conclusions 10%

Groups

Assignments are submitted by groups of of 1 to 3 students. Different elements may have different grades. Other group sizes will not be considered.

It is advisable that the students from the same group perform overlapping work and only after that, exchange ideas with each other. Group work is important for learning from other people.

Submissions

Formal final deadline is **January 14th 2022**, to be submitted in moodle, and only in moodle. Submissions after that date will be multiplied by a monotonously decreasing factor that starts in 1.

- Intermediate submissions:
  - **December 3**: html version of the notebook with the first 2 CRISP-DM phases and part of the 3rd.
  - **December 22**: html version of the notebook with the work including modeling or beyond.

Ethical principles

When submitting, students commit themselves to follow strong ethical principles. All the work must be done by the elements of the group alone. All members of the group will be involved with the whole of the work. All the materials used and consulted must be credited in the work.