Machine Learning - Project 2020

Predicting online purchasing intention

Souhaib BEN TAIEB

14 April 2020

Task

The purpose of this project is to produce on the basis of various predictors the best model to predict if a session on an e-commerce website will lead to a purchase. This can be used for example to offer specific content only to those who intend to purchase and not to offer content to the other users. The competition (with related datasets) is hosted on the following Kaggle website: https://www.kaggle.com/t/4acd5a9fcf3f402cb76b2a6d35c727fe.

The dataset has been split into training and test sets. The full training set is available to you. But only the features are provided for the test set. You can evaluate your predictions by submitting them to the Kaggle website. Note that only a random subset containing 55% of the test set is used to compute your *public score*. Your final *private score* using the full test set will be provided at the end of the competition.

The dataset consists of 17 features associated to 7,396 sessions on the e-commerce website. Each session belongs to a different user in a 1-year period. The 'Transaction' feature is the output variable, which indicates wheter the session has been finalized with a transaction. The other features are described below.

"Category" and "Category_Duration" represent the number of different pages (in a certain category) visited in that session and total time spent in this category, respectively. These values have been extracted from the URL of the pages visited by the user and updated in real time when a user moves from one page to another. There are three categories related to "account management" (Category I), "Web site, communication and address information" (Category II), and ""product related" (Category III).

The "Bounce Rate", "Exit Rate" and "Page Value" features are metrics measured by "Google Analytics" for each page in the e-commerce site.

The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day). For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

Here are some additional features: - OS: Operating system of the visitor. - Browser: Browser of the visitor. - Region: Geographic region from which the session has been started by the visitor. - TrafficType: Traffic source by which the visitor has arrived at the Web site. - VisitorType: "New Visitor", "Returning Visitor" and "Other". - Weekend: Boolean value indicating whether the date of the visit is weekend. - Month: Month value of the visit date.

The evaluation metric for this competition is the F1-score (https://en.wikipedia.org/wiki/F1_score).

1. Your first task is to form a team with two people.
2. Each team member should create a Kaggle account (using your UMONS email address)
3. Form a team on Kaggle.
4. Do some basic exploration of the dataset
5. Build your first model. Predict the test set, and upload your predictions to Kaggle.

6. Try, and try again to improve your model. You can make a maximum of five submissions per day.

Project report

The data analysis report can be a maximum of **10 pages**, and must abide by the section structure described below.

- Section 1: Introduction

The introduction will describe the data set and motivate the problem. It should be brief.

- Section 2: Methodology

This section describes the models and methods you have used, including a justification of your choices. You should also present your model fitting, diagnostics, etc.

You should discuss and compare at least three different classification models.

- Section 3: Results and Discussion

This includes for example graphs and tables, as well as a discussion of the results.

- Section 4: Conclusion

This includes summary of the findings.

You should clearly explain what you have done, using figures to supplement your explanation. Your figures must be of proper size with labeled, readable axes. In general, you should take pride in making your report readable and clear. You will be graded both on *machine learning content* and *quality of presentation*.

Grading

- Total points: 20
- Accuracy of classifier on Kaggle: 6
- Report: 14

Deadlines

Do not wait until the last minute. Late submissions will not be allowed.

- April 22, 11:55pm: Submit the names of each member of the team on Moodle.
- May 3, 11:55pm: At least one Kaggle submission needs to have been made.
- May 17, 11:55pm: The Kaggle competition closes.
- May. 24, 11:55pm: Upload to Moodle your project report and code, one per group.