Machine Learning - Project 2020 report

Predicting online purchasing intention

Florent HUYLENBROECK
Laurent BOSSART

May 24, 2020

Introduction

The goal of this project was 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 project took the form of a competition hosted on `https://www.kaggle.com`.
In groups of two, we had been given two dataset, one for training and one for testing, a short explaination of the different predictors and a way to measure our model's efficiency (the *F-1 score*).
We were able to submit five model per day.
This report will describe our reflexions on the subject and what submission we made.

The data sets

The predictors used in the two datasets are the following :

- *CategoryN* and *CategoryN_ Duration* with $N \in \{I, II, III\}$ represent the number of different pages belonging to a certain category visited by the user during that session and the time spent in that category.
  I = account management pages.
  II = website, communication and address information pages.
  III = product related page.

- *Bounce rate*, *Exit rate* and *Page value* are metrics provided by "Google Analytics" for each pages in e-commerce.
  Bounce rate is the number of single pages viewed by user (meaning the user exits the website on the same page he entered it, without navigating the site further).
  Exit rate tells from which page the users exit the website the most.
  Page value is the number of pages that a user visited before completing a transaction.

- *SpecialDay*, *Weekend* and *Month* all give information on the date when the session started.
  SpecialDay indicates the closeness of the site visiting time to a special day.
  Weekend tells if the session started during a saturday or a sunday. *Month* is the month of the visit date.

- *OS* and *Browser* are the exploitation system and the browser used by the user.

- *Region* is the geographic region where the user started his session.

- *TrafficType* is the traffic source from which the user entered the website.

- *VisitorType* indicates whether if the user is returning or new.

- *Transaction* indicates if a transaction has been completed. It is the value we will try to predict on our models.

Methodology

Brainstorming

Before anything else, we tried to think logically about the predictors. We ordered them from most to least important. We came up with a list that helped us build our first naive models.

Crossvalidation and useful functions

The first thing we did in *R* was to implement various function that would made our experimentation easier. We so implemented three functions :

- **submit_ prediticion**(*model*) that, given the parameter *model* being a anonymous function, returns a model, use it to predict our testing set's *Transaction* value and write that prediction next to the matching *Id*'s in a .csv file for submitting on Kaggle.
- **f1_ score**(*prediction*) that given a prediction over the training set, evaluates the $F1$-*score* of that prediction.
- **crossvalid**(*model, nrep, pring*). This functions performs a 10-fold cross validation of the model *model* a number *nrep* of time over the training set and returns the mean $F1$-*score*. The argument *print* serves a debugging purpose.

Data pre-processing

Results and Discussion

Conclusion