Stefano Samele stefano.samele@polimi.it Matteo Matteucci matteo.matteucci.polimi.it

Machine Learning Course 2023

Homework

In this project, we aim to use the Online Shoppers Purchasing Intention dataset to build a classification model that can predict whether or not an online shopper has the intention to make a purchase.

Background

The Online Shoppers Purchasing Intention dataset is a publicly available dataset that contains information on the browsing behavior and purchase history of online shoppers. The dataset contains different sessions, and it was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The dataset has been made available on the UCI Machine Learning Repository.

Dataset Composition

The dataset contains just over 12,000 samples and includes both categorical and continuous variables. The categorical variables include information on the type of the operating system, the region of the shopper, and whether or not the shopper was a returning visitor. The continuous variables include information on the duration of the visit, the number of pages visited, and the exit rate, among others.

We have provided you with a particular version of this dataset. There are two files representing the training set and the test set. For the purposes of this project, we have corrupted a significant amount of samples of the Exit Rate variable.

Below you can find the description of all the dataset attributes.

- *Administrative:* number of pages of the website related to the administrative section that the user visited.

- *Administrative Duration:* total time spent by the user on the website's administrative pages.
- *Informational:* number of pages of the website related to the informational section that the user visited.
- *Informational Duration:* total time spent by the user on the website's informational pages.
- *Product Related:* number of pages of the website related to the product that the user visited.
- *Product Related Duration:* total time spent by the user on the website's product pages.
- *Bounce Rate:* percentage of visitors who enter the website and leave without viewing any other pages during that session.
- *Exit Rate:* percentage of visitors who leave the website from that page during that session.
- *Page Value:* average value of the page that the user visited before landing on the goal page during that session.
- *Special Day:* closeness of the visit to a special day such as Mother's Day or Valentine's day.
- *Month:* month of the year in which the visit took place.
- *Operating System:* type of operating system used by the user.
- *Browser:* type of browser used by the user.
- *Region:* region of the user.
- *Traffic Type:* type of traffic through which the user reached the website.
- *Visitor Type:* type of visitor, i.e., new or returning.
- *Weekend:* categorical feature representing whether the visit took place on a weekend or not.
- *Revenue:* This is the target variable, a binary feature that indicates whether the visitor made a purchase or not.

Project Steps and Requests

1. Perform preliminary analysis of the data. For instance, but not limited to visualizing samples, identifying if features are correlated, determining which are most correlated with the target class, and inspecting the distribution of samples among classes. We want to recover from the data loss of the Exit Rate variable. Train a regressor to predict the values of this variable, given the remaining ones. Compare different regression algorithms for this task, and perform features selection. Since this feature is also missing in the test set, use only the training data for this step and use robust evaluation techniques to compare algorithms.

2. Use the regression model trained at the previous step to recover Exit Rate by predicting their value for each missing sample on the train and test set. Build a **classification model** to correctly predict if an online shopper is likely to perform a purchase. Perform features selection, compare different algorithms, and identify the one that works the best on this dataset. Finally, test the performance of the best algorithm on the provided test set. Determine whether a model built using the Exit Rate information is better than a model built using the remaining features.

3. Repeat the analysis done at step 2 with clustering based algorithms; compare the performance with respect to the best classification model.

Submission

The project deadline is 31 July at 23:59 CEST (Rome)

- Create a Jupyter notebook to answer all the requests, using the libraries presented during the laboratory classes.

- Include Name, Surname and Student ID in the notebook.

- You are free to use the structure that you prefer within the notebook. However, please use markdown cells (Cell > Cell Type > Markdown) to insert section titles and clearly identify the different requests. You are free to add subsections to make the notebook more readable.

- Add text cells (markdown) to briefly explain what you did and why, and to help you answer the requests.

- Please, check that the notebook can execute correctly before submitting your work.
  Press Kernel > Restart & Run All and check that all cells execute correctly without errors.

- The output of notebook cells should be included in the submitted notebook, i.e. the notebook should be submitted as already executed.

Evaluation

Evaluation

The project evaluation is not based on the regression and classification scores you obtain in each request but rather on the soundness of your analysis and the choices you made to solve issues (if any) you may encounter. A good project follows the evaluations discussed during the lectures and lab sessions while adapting and integrating the analysis based on the project domain.

Points will be given for each step of the project according to the following table:

0	Completely wrong, or one key concept is not taken into account
1	Ok, but poorely discussed. E.g., just a copy-paste of labs with no comments, or some conceptual errors
2	Well done, good comments and analysis, adapted code/analysis, he/she interpreted the results

Penalty points will be subtracted according to the following table:

0	Does execute with no errors
-1	Does not execute, but it is an easy fix to make it work
-2	Does not execute, and it requires more that 10 mins of debugging the errors