Assessment Task for Applicants at MFG Data Science Team

MFG
MANAGEMENT FINANCIAL GROUP

Contents

1. Task Description

Please examine the dataset with Google Analytics data of online shop visitors provided on the following link:
https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

The target variable “Revenue” labels the visitors of the online shop as clients that have made a purchase (true) and clients that haven’t completed a purchase (false).

Please tune, test and evaluate with the appropriate metrics at least two machine learning classification algorithms for the purpose of creating a predictive model that generalizes well on new clients predicting whether a client is going to make a purchase from a website or not based on the explanatory variables provided in the dataset. Please comment on the results and select the one best performing algorithm for the task at hand. Comment on the possible business applications of the final predictive model selected.

This is a high-level description of the features:

- **Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration** - number of different types of pages visited by the visitor in that session and total time spent in each of these page categories

- **Bounce Rate** - Google Analytics Metric. The percentage of visitors who enter the site from that page and then leave (“bounce”) without triggering any other requests to the analytics server during that session

- **Exit Rate** - Google Analytics Metric. Feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session

- **Page Value** - the average value for a web page that a user visited before completing an e-commerce transaction.

- **Special Day** - indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine’s Day) in which the sessions are more likely to be finalized with transaction.

- **Operating system** - Operating System of the user.

- **Browser** - Browser of the user.

- **Region** - Region of the user
- **Month** - month of the transaction
- **Weekend** - Flag if the transactions was made during the weekend.
- **Traffic type** - Traffic Type of the user
- **Visitor type** - New or Returning Visitor
- **Revenue** - The target column. TRUE values mean a purchase was made and FALSE mean a purchase was not made. This is the target variable we are going to build a classification model to forecast.

2. Expected Deliverables

- A report in word, pdf, Markdown or Notebook format with the results of the analysis.
- A script in the programming language used for the analysis that can be run and used in order to reproduce the results.

3. Preferred Tools for Implementation

R or Python Programming Language.

4. Tips

- Provide a section in the report with task and data description.
- Provide a section with exploratory data analysis.
- Divide the data in train and test samples to ensure a robust performance on new data.
- Provide a section with algorithms testing and evaluation.
- Test at least one simple algorithm like Logistic Regression and at least one more complex algorithm like Random Forrest.
- Provide a section with the final results, final model selection and comments on possible added value for the business.

5. Expected delivery time

5 calendar days.