Revenue Prediction in Online Advertising

MATH2319 – Machine Learning, 2019

Dataset Information

This dataset contains real-world online advertising data where the target feature is a revenue-related metric and the descriptive features are various advertising metrics and characteristics. Each row represents a website traffic record that comes from a specific country, company, and device type combination. The dataset contains 30 days of training data and 5 days of test data. The training data contains about 215K records and the test data contains about 31K records.

Features in this data set are as follows:

- **companyId:** Company ID of record (categorical)
- **countryId:** Country ID of record (categorical)
- **deviceType:** Device type of record (categorical corresponding to desktop, mobile, tablet)
- **day:** Day of record (integer between 1 (oldest) and 30 for train, 31 and 35 (most recent) for test)
- **dow:** Day of week of the record (categorical)
- **price1, price2, price3:** Price combination for the record set by the company (numeric)
- **ad_area:** area of advertisement (numeric)
- **ad_ratio:** ratio of advertisement's length to its width (numeric)
- **requests, impression, cpc, ctr, viewability:** Various metrics related to the record (numeric)
- **ratio1, …, ratio5:** Ratio characteristics related to the record (numeric)
- **y (target feature):** revenue-related metric (numeric)

Problem:

The problem at hand is to predict the value of y for each website traffic record in the test data. Even though the dataset contains a day value, you must treat the day feature (or any new features you derive from the day and/ or dow features) as just another input to your algorithm and avoid using time series/ forecasting methods in your solution.

Please feel free to define new descriptive features based on the existing ones. Once you have pre-processed the data, you are free to apply whatever machine learning (ML) algorithm you like (such as linear regression, support vector regression, polynomial regression, neural networks, deep learning, etc). Programming language is Python 3 and there are no library restrictions.

Conditions of Use

This dataset is exclusively for the use of MATH2319 students. You must not share this dataset with anyone else outside of this course and you must not post it in any public domain in any way. This is a strict requirement.

Instructions for Course Project Only:

You must use only the training data for both phases. In this case, you will simply ignore the test dataset and just use cross-validation RMSE (root mean square error) on the training dataset as your performance metric. As usual, whatever ML algorithms you use, you must specify their parameters in your phase 2 project report.

Instructions for Kaggle Only/ Kaggle + Project:

Please use only training dataset for phase 1 and, both the training and test datasets for phase 2. As usual, whatever ML algorithms you use, you must specify their parameters in your phase 2 project report. In case of Kaggle only, we will request from a few of the top performing teams that they share with us their Python code, which we will share with the company that kindly provided this dataset for our students.

Please keep in mind that the training dataset has all the y values whereas the test dataset does not contain any y values. We actually know these y values, but we are apparently hiding them from you so that we can figure out the best team on the Kaggle competition!