

Conversion Prediction Write Up

Background

Analytics is a powerful tool for understanding customer behavior. Most e-commerce sites track a number of metrics when users interact with their site - the number of searches, number of product clicks, number of cart adds, etc. Most analytics engines group these metrics into sessions. A session is simply a visit to and interaction with a website (generally, sessions end when a user leaves the website or is inactive for a long period of time). One of the most important metrics to consider for e-commerce is conversion rate - the percentage of sessions that result in a sale. If we can discover patterns in conversion rate, we might be able to make the shopping experience more enjoyable for customers.

Problem Description

You are tasked with developing a model that can accurately predict whether a session results in a sale.

Training Data

Training data is an array of JSON objects. Each object stores metrics describing a session. These are the fields in each object:

id: unique integer value that identifies the session

startLocalDateTimeString: a string representing the date and time when the session began, relative to the timezone the session's user resides in

startUnixTimeSecondsGMT: number of seconds since the unix epoch when the session began

endUnixTimeSecondsGMT: number of seconds since the unix epoch when the session ended

numClickstreams: number of clickstream events in the session (a clickstream event is a generic action performed)

numSearches: number of times a search occurred in the session

numNoResultsFound: number of times a search returned no results

numAutoCompleteClicks: number of times an autocomplete suggestion was clicked in the session

numRelatedTermClicks: number of times a "related query" suggestion was clicked in the session

numSuggestedProductClicks: number of times a product suggestion was clicked in the session

isMobileVisit: boolean denoting whether the session occurred on a mobile device

isSale: boolean denoting whether the session resulted in a sale

Example training data object:

```
{ "id":8890,"startLocalDateTimeString":"2017-06-07T14:48:01","startUnixTimeSecondsGMT":1496861281,"endUnixTimeSecondsGMT":1496862791,"numClickstreams":37,"numSearches":5,"numNoResultsFound":0,"numAutoCompleteClicks":1,"numRelatedTermClicks":0,"numSuggestedProductClicks":0,"isMobileVisit":true,"isSale":false}
```

Testing

Your model should read in the test dataset and output a CSV file with two fields (do not include a header): **id** and **isSale**.

Input

The test dataset has the same exact structure as the training dataset *except* for the **isSale** field - the **isSale** field is not included in the test dataset. Also note that while the structure is the same, the actual sessions are different. No sessions that appear in training data are present in testing data. Also note that session IDs are only unique with respect to their dataset - that is, the session with **id=1** in the training data is different than the session with **id=1** in the testing data.

Example testing data object:

```
{ "id":56,"startLocalDateTimeString":"2017-06-09T08:55:56","startUnixTimeSecondsGMT":1497012956,"endUnixTimeSecondsGMT":1497016190,"numClickstreams":41,"numSearches":4,"numNoResultsFound":0,"numAutoCompleteClicks":0,"numRelatedTermClicks":0,"numSuggestedProductClicks":0,"isMobileVisit":false}
```

Output

Your model should output a CSV file with two columns and no header row. The columns should be:

`id, isSale`

Where `id` is the id of a session object in the testing data, and `isSale` is a boolean representing whether your model predicted the session with that id resulted in a sale.

Example output:

`123456, true`

`987654, false`

`102030, false`

Scoring

Your model will be scored by it's prediction accuracy: (number of correct predictions) / (number of total sessions).