Exercise: Conversion Rate

Goal

You are working for an e-commerce company. The company wants to optimize the conversion rate for its website. Conversion refers to the action the company wants the user to take, which in this case is making a purchase. The conversion rate is simply the number of conversions divided by the total number of visitors the page receives.

Optimizing conversion rates is a very common task for a data scientist, and rightfully so. We are able to collect all sorts of data about people who buy something on our website as well as about people who don't. This gives us a tremendous opportunity to understand what's working well (and potentially scale it even further) and what's not working well (and fix it).

In brief, the goal of this exercise is to explore the available data, build a model that predicts conversion rate and, based on the model, come up with ideas to improve revenue.

The problem is very straightforward, there are no dates, no tables to join, no feature engineering required. This exercise is a great starting point to get familiar with the general structure of data challenges, which are a common part of interviews for Data Scientist roles.

Exercise Description

We have data about users who hit our site. We know whether they converted or not, as well as some of their basic characteristics such as the country they're in, the marketing channel that brought them to the site, their age, whether they are repeat users, and the number of pages visited during that session (as a proxy for site activity/time spent on the site).

Your project is to:

1. Perform Exploratory Data Analysis and produce 2-3 plots that show either important features or interesting patterns in the data. It is up to you what you want to highlight.
2. Build a model to predict conversion rate and critically evaluate it, explaining your choice of model and performance metric.
3. Come up with recommendations for the product team and the marketing team to improve conversion rate.

Deliverable

The deliverable for this exercise consists of a report both in the form Jupyter notebook file (use Python!) and a PDF of that report. Your files should be named in the following format:

- "FIRSTNAME_LASTNAME_presession_exercise.ipynb"
- "FIRSTNAME_LASTNAME_presession_exercise.pdf"

substituting your name for the appropriate placeholder. The files should be pushed to a private github repo (instructions) named "FIRSTNAME_LASTNAME_presession_exercise") that has the Program Directors added as collaborators (instructions, please check the original email for our Github IDs). Once you've completed the exercise, please fill out this survey. The survey contains a field where you can provide a link to the private repo that contains your solution files.

Reports like this are a common way of sharing your work as a Data Scientist. Since these reports will often be read by people who are less familiar with your work than you are, it is crucial to both 1) comment your code and 2) explain the analysis/modeling steps you're taking and why (Markdown!).

While the former is most important for your future self and other people who might have to work with your code later on (both of whom will thank you, trust us!), the latter is important for anyone who is reading your report and wants to evaluate the work you've done and the conclusions you're drawing from it.

Data

The data for this exercise consists of 1 table which you can download by clicking here.

The table is called "conversion_data". It has information about signed-in users during one session. Each row is an individual user session.

Columns:

- **country** - user country based on the IP address
- **age** - user age. Self-reported at sign-in step.
- **new_user** - whether the user created the account during this session or had already an account and simply came back to the site
- **source -** marketing channel source
    - **Ads** - came to the site by clicking on an advertisement
    - **Seo** - came to the site by clicking on search results
    - **Direct** - came to the site by directly typing the URL on the browser
- **total_pages_visited** - number of total pages visited during the session. This is a proxy for time spent on site and engagement during the session.

- **converted** - this is our label. 1 means they converted within the session, 0 means they left without buying anything. The company goal is to increase conversion rate: # conversions / total sessions.

Example

Let's look at the characteristics of the user in the first row.

Field	Value	Description
country	UK	The user is based in the UK
age	25	The user is 25 yr old
new_user	1	The user created their account during this session
source	Ads	The user came to the site by clicking on an ad
total_pages_visited	1	The user visited just 1 page during that session
converted	0	The user did not buy during this session. These are the users whose behavior we want to change!