Data Science Internship 2023

Anonymised

Context Enhanced Audience Modelling

About the Company

Anonymised is a data privacy startup based in London and Barcelona. We are creating a more private and secure web by using machine learning for data anonymisation. Digital marketing requires companies to collect data about their users and combine it with other data to identify the individual, understand their preference and deliver targeted advertising. But with the development of privacy regulations around the world, this trade in personal data is no longer lawful. Anonymised allows marketers to find consumers in a privacy-friendly way, by storing all personal data on the device and anonymising it with machine learning.

About the internship

We are looking for ambitious, confident and curious data scientists to work with us. You will be part of a young team of developers pushing the boundaries of innovation in data science. At the moment, we have one internship project available. We expect the internship to start in March 2024, and to last 4-6 months. The internship project can also be used in your dissertation, and we will support you taking some time off at the end of the project to write up the results. We offer a reimbursement of €800 per month, and we will expect you to work full time (except for time allowed spent writing the dissertation). The internship is fully remote.

About the Project - Context Enhanced Audience Modelling

Targeting users with correct advertisements is done through analysing user browsing habits and behaviours. However this data is not always available and we frequently need other data to help us target the right audiences online.

The goal of this project is to extend Anonimised's privacy-preserving targeting capabilities by utilising ML-driven probabilistic models ("lookalikes") to enrich and expand first-party (i.e. single-domain) audiences. As an example:

- User A (seed audience) has known attributes ABCD-EFGH
- User B (seed audience) has known attributes ABCD-GHJK
- User C (lookalike audience) has known attributes ABCD only
- When can confidently predict that User C is also likely to have GH attributes?

To build lookalike audiences we need a technique that computes the similarity among users with known attributes (seed audience) and compares it with users with fewer attributes (unknown audience) to probabilistically determine which unknown users should be treated as seed audiences (lookalikes).

The project will include several challenges:

- You will help us identify and expand the array of attributes that can be used to calculate similarity between seed and lookalike audiences.
- You will design a modelling technique that helps us define an audience that is present on a certain domain.
- You will need to analyse low-resource users and how they behave on the domains we are present on, which will include statistical analysis on the correlation between different topics they browse and which ad campaigns they respond to positively.
- You will experiment with different state-of-the-art classification models looking for the algorithm that provides the best results. The performance will be measured using different metrics including Click-through rates and conversion.
- You will get acquainted with a range of Big Data products offered by the Google Cloud Platform (such as BigQuery, BigTable, Dataflow, Dataform, etc.)
- You will learn how to build end-to-end automated pipelines for the whole process using different MLOps tools (Apache Beam, Apache Airflow, MLflow).

Contact:

1. Danail Krzhalovski : danail@anonymised.io
2. Giovanni Vedana : giovanni@anonymised.io