Dataset

https://developer.ibm.com/data/bias-in-advertising/

https://towardsdatascience.com/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285

Report

- **Emily**: What did you expect before conducting your analysis? Did you have any hypotheses around the problem, such as features that you expected to be important or unimportant?
- **Naila:** Why did you select the analysis tools that you used?
- **Naila:** What did you learn from your analysis?
- **Jacob**: To what extent do you believe your findings are generalizable to other, similar contexts (e.g. for Telco churn, to customer churn in general)? Why?
- **Jacob**: What is your proposed solution and why did you select it?
- **Jacob**: Is your proposed solution fair? Why or why not? How can you tell?

Ethan: Distribution, bar charts, and documentation of the Jupyter Notebook.

Team Members
Emily Do
Naila Hajiyeva
Jacob Salazar
Ethan Wen

Timeline:

- Meet every Friday 5-7PM for check-in and updates
- Complete report by 11/28
- Presentation creation + practice 11/28-12/05

Chosen Topic
Our team has chosen to study the presence of bias in advertising, particularly how it is used in converting their target audience, with the conversion being defined as users interacting with the ad. We will examine how advertisements are targeted for certain users based on user information that could have been collected from them, such as age and gender. As such, we are interested to see if any particular attributes about users have the most success in conversion prediction.

This study is interesting to us because it can not only be used by stakeholders in the advertising industry, but also bring awareness of bias in the content that consumers like us are being exposed to. This could bring a stronger sense of consumer autonomy in deciding what kind of brands and products they are interested in purchasing. In addition, this can tie in to a larger discussion of consumer consent to sharing information about themselves and allowing data collection.

Data
Our team has chosen to study an IBM dataset "Bias in Advertising" which is made up of synthetically generated data for users who were shown a particular ad. Each data point represents a specific user, along with their attributes including gender, age, income, political/religious affiliation, parental status, home ownership, area (rural/urban), and education status. This also comes with data on user conversion, which describes the ad clicked. This included three variables of the predicted probability of conversion, the predicted conversion (binary 0/1), and the true conversion (binary 0/1) of if the user actually clicked on the ad or not.

Plan:
Analysis Objectives:

- ~~Clean the data from unknown values for gender, income, and area~~
- Feature importance in predicting the ~~probability of conversion~~
- Feature importance in predicting true conversion ~~- summary statistics~~
    a. ~~What has a higher weight in terms of true conversion?~~
    b. ~~Calculate false positives and false negatives in true conversion and probability of conversion.~~
- ~~Overview statistics (kernel density graph)~~
- ~~Removing records with unknowns (ex: age, gender). → smaller subset of data with more features~~
- College Educated, Gender, Age, Homeowner, Income

Expected results:
We expect the following features (gender, age, income, college-educated) to have a higher feature importance in terms of conversion prediction.