Spencer Tollefson

October 17, 2018

Proposal for Project McNulty

Projecting Smoking Habits Based on OkCupid Profiles

OkCupid.com is an online platform primarily known for online dating, while also providing services for friendship, social networking, and "specialty" dating such as for conservatives and for vegans. The site was launched in 2004 and as early as 2007 had some of the best name recognition among its competitors.

For a monthly fee - varying based on the amount of features a user desires - customers create a profile that is meant to illustrate unique and interesting facts and interests. Some characteristics are descriptive and straightforward, such as age, height, and sex. Others are similar, but are topics considered more private or vulnerable such as income, education level, sexual orientation, and religion. Then there are text fields which ask questions, allowing customers to express themselves creatively with their answers and also indicate what they are looking for in their partners.

Once their profile is created, customers are able to browser other profiles, send messages, and hopefully use the platform to meet and make plans to date compatible individuals.

Question/Need:

Predict if an individual identifies as someone who smokes tobacco based on their other OkCupid profile attributes.

Many people demonstrate they are willing to share information online that they would not share in person to a physician, family, or friends. Operating under the assumption that some people are willing to share knowledge of their social "vices" in their dating profile, but not with their doctors or concerned family and friends, a "smoker" prediction model would aid people in deciding if someone they know smokes.

Description of data:

The data is a table of OkCupid users' attributes pulled from 2012 in the San Francisco region. OkCupid provided the anonymized dataset to Albert Y. Kim and Adriana Escobedo-Land for use in their research published in the Journal of Statistics Education, Volume 23, 2015.

The Github can be found here.

This dataset was released on the heels of a dataset published in 2,014 by Danish researchers. That data set was **not** anonymized and its release sparked backlash against the researchers as it exposed details of 68,000 individuals publicly.

Features	Type	Number Categories	Planned number categories	Description	Plan to Use in Model	Availability for Profiles
age	integer	NA	NA	Age in years	Y	All
body_type	string	13	8	Description of body dimensions	Y	Nearly all
education	string	33	8	Level of education completed	Y	Nearly all
height	integer	NA	NA	Height in inches	Y	All
income	integer	13	8	Income bracket / tier	Y	All
job	string	22	22	Job title	Y	Most
orientation	string	3	3	Sexual orientation	Y	All
pets	string	16	3	Likes and ownership of cats & dogs	Y	Over half
sex	string	2	2	Male or female	Y	All
smokes	string	6	2	Description of smoking freq	Y	Nearly all
status	string	5	5	Current relationship status	Y	All
drinks	string	7	NA	Description of drinking freq	N	Nearly all
ethnicity	string	218	NA	Description of ethnicity	N	Most
offspring	string	16	2	Viewpoints/status of having kids	N	Less than half
drugs	string	4	2	Description of drug use freq	N	Most
diet	string	19	NA	Description of dietary preferences	N	Over half
religion	string	46	3	Religious beliefs and intensity	N	Over half
speaks	string	NA	NA	Combo of languages spoken and how well	N	Nearly all
location	string	199	NA	Location by municipality/neighborhood	N	All

Features	Type	Number Categories	Planned number categories	Description	Plan to Use in Model	Availability for Profiles
last_online	DateTime	NA	NA	DateTime indicating last time on OkCupid	N	All
essay(0-9)	string	NA	NA	Text field replies to questions	N	Many missing

Characteristics of each row of data:

Each row of table has fields of information pertaining to one individual. At the earliest stages of EDA, I decided to drop the text response fields from each application as there will be NLP component for this project.

Known Unknowns

- All information in this table is self-reported. As such, it is likely that some users will untruthfully answer questions. Particularly, topics socially considered taboo such as drinking, drug use, and smoking are strong candidates for eliciting untruthful answers. It is unknown which responses are untruthful and how many there are.
- Although OkCupid.com procured this dataset, it is unconfirmed if the profiles are authentic.
- Drug use, drinking, and tobacco smoking habits potentially may have high correlation.