Project McNulty - Online Dating: Who Wants Children?

Spencer Tollefson

October 31, 2018

Project Design

OkCupid.com is an online dating platform catered toward single adults who are seeking romantic partners. 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.

Upon getting my hands on a dataset of nearly 60,000 OkCupid.com profile attributes, I became interested if the other attributes were indicative of people's desire to want or not want children. Interestingly enough, the dataset showed that many people chose not to provide this optional information. My assumption was this is because people do not want to be instantly filtered as a "no" if their children rearing plans do not match people who may well be good potential matches.

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.

A model that performs well in predicting one's desire to raise children would have multiple uses. Potential uses for a high performing model include, but are not limited to, the following:

- Hosted as a web app where people may input information of someone they are interested in. The web app would output the model's predicted probability that the individual wants to have children.
- Improve matchmaking capabilities of dating services.
- Essentially a misuse, but for health insurance companies to decide how likely it is their customers will have children in the future.
- Another misuse. HR departments of various companies to use to determine how likely it is a potential or current employee will have a child and miss work time.

Tools

- Python:
    - Data cleaning: Pandas, Joblib
    - Data analysis: Numpy, Pandas, Scikit-learn, Jupyter
    - Presentation: Matplotlib, Seaborn
- Google Slides

Data

The data was obtained from a Github repo. The original data was scraped via a Python script in June 2012. It contains profile attributes of nearly 60,000 OkCupid.com accounts which had been active during the year preceding June 30, 2012, have a profile picture, and have their city listed as within a 25 mile radius of San Francisco, CA.

All features may be found in the appendix. Cleaning was applied to many columns to account for grouping of multiple subjects. This in effect became the only form of feature engineering I applied to the data. For example, the original data grouped religious intensity (serious, not serious, no care at all) and the name of the religion together. By cleaning I separated these two subjects and considered them separate features.

Ethical questions bombarded me as I dealt with this data set and the entire project. Is it ethical to create a model to help people "screen out" others while dating? How about using ethnicity, religious, and sexual orientation as features for deciding if someone wants to have children? How about the people whose profiles were used as the data observations?

While I did not delve too deeply into the implications due to time constraints and the likelihood that this project would not go viral, it did make me think about this data.

Results

I created models of 6 different classifier algorithms. All used sk-learn library tools and typically used the GridSearchCV or RandomizedSearchCV to hone in on some of the best tuning parameters. The gradient boosting implementation performed best, with an AUC ROC of 0.774. The others all fell in between that value and 0.720; thus a fairly narrow range of values.

What I would do differently

Age was the feature of strongest correlation to the target variable. See if certain ages have strong correlations with the target (i.e. perhaps under 21 above 40 years old definitively do not want children) and filter out observations with those features to make the modeling more robust on a harder to nail-down age group.

Assign ordinal values to certain features. For example, intuitively it appears there is a ranking to be derived to how much someone consumes alcohol when the optional descriptors "rarely", "socially", and "often" are used. There are numerous categorical features which I believe ordinal values could have been derived and modeled.

There was a "maybe" class for the target variable of "want offspring". With more time, I would have made a multi-class model that attempted to classify profiles as "no", "maybe", or "yes" for the target variable of wanting children.

Consider the attributes of people who **did** choose to list their children preferences to those who **did not**. This would give a better idea of how well the model could be extrapolated to a large proportion of the population which chose not to disclose that information.

Think more carefully about the ethical considerations involved in which features are used in the modeling. I purposely removed ethnicity and income level as I did not want my model to discriminate, but used features such as sexual orientation and religion. Although all of this information was freely offered by users in making their profiles, times are changing as people become more aware of how their data is and can be used. Privacy preferences are being updated as well.

Considered more deeply about the real-world value of this predictive model.

Appendix

Dataset Features

Features	Type	Number Categories	Planned number categories	Description	Used in Model	Availability for Profiles
age	integer	NA	NA	Age in years	Y	All
body_type	string	13	8	Description of body dimensions	N	Nearly all
education	string	33	8	Level of education completed	N	Nearly all
height	integer	NA	NA	Height in inches	N	All
income	integer	13	8	Income bracket / tier	N	All
job	string	22	22	Job title	N	Most
orientation	string	3	3	Sexual orientation	Y	All
pets	string	16	3	Likes and ownership of cats & dogs	N	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	Y	Nearly all
ethnicity	string	218	NA	Description of ethnicity	N	Most
offspring	string	16	2	Viewpoints/status of having kids	Target	Less than half
drugs	string	4	2	Description of drug use freq	Y	Most

Features	Type	Number Categories	Planned number categories	Description	Used in Model	Availability for Profiles
diet	string	19	NA	Description of dietary preferences	N	Over half
religion	string	46	3	Religious beliefs and intensity	Y	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
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