Mental Health in Technology – Data Quality Plan

Miriam Callahan

NOTE: The original data set can be found at https://data.world/quanticdata/mental-health-in-tech-survey. All changes have been accounted for, but the omitted data has not been placed here.

Part I – Deletion & Changes of Dataset Values

- Removal of Timestamps
    - I decided to remove all the timestamps from the data set as all respondents answered within a timeframe of two days (August 27-28, 2014) so I do not find it statistically significant since there was not a massive time difference between responses.
- Deleting Respondents with Erroneous Ages
    - Upon first glance of the dataset, I noticed there were a handful of respondents who put unrealistic ages within the data. Putting such information suggests they were not taking the survey seriously, hence I deleted them to keep only legitimate and credible responses in the data. The following respondents and the ages they put (along with other reasons for deletion) have been put below:
        - Respondent #366 – Respondent #366 was omitted for putting his age as 329.
        - Respondent #717 – Respondent #717 was omitted for providing an age of -1726.
        - Respondent #736 – This respondent's age was listed as "5." While this have may been a typo, there are many possibilities for this value so for the sake of time, it was deleted.
        - Respondent #991 – Respondent #991 was omitted for providing an age of 8 (an age where someone is unlikely to hold a job), put in gender as "A little about you," country as "The Bahamas" but their state as Illinois. All of this contradictory information suggests the surveyor did not take the survey seriously.
        - Respondent #1092 – Respondent #1092 was omitted for providing an age of 11 where they are unlikely to hold a job.
        - Respondent #1129 – Respondent #1129 was omitted for providing an age of -1 and their gender of "p."
- Changing Respondents with Erroneous Ages
    - Respondent #145 put his age in as -29. I assume this was a mistake, so I changed it to 29 instead.
- Deleting Comments
    - Since comments were so varied to be of use when predicting a nominal value, I deleted them from the data set to make analysis easier.
- Deleting Countries Other Than The US

    - Although this was a worldwide study, 746 of the 1253 respondents were American. In many cases, international respondents were the only ones representing their nation causing an issue of severe underrepresentation. However, since most of the time only 1 respondent was representing a singular nation changing proportions of representation of the data may be problematic as it does not represent the diversity of those nations. Similarly, many countries are not represented at all. As a result, all respondents who indicated they were from outside the US were deleted and the country feature as well. The table provided in the DQR demonstrates this point.
- Deleting Respondents with Non-Existent Genders
    - Other than the aforementioned respondents, some other respondents were deleted for putting in genders which insinuated they did not care about the survey.
        - Respondent #388 was omitted for putting their gender as "Nah."
- Normalizing Genders
    - In order to have a more streamlined view of gender, I decided to normalize similar responses of gender such as M, mail and male to Male. Since an LGBTQ+ identity (particularly identifying as trans within the discussion of gender) can have a great impact on one's mental health, I divided the classifications to include both cisgender and transgender individuals. Should an individual have entered "trans" or some variation thereof into the gender category, they would be placed in the trans category of their corresponding gender. Should such an identifier not be present, they would be placed in the cisgender category of their respective gender. The following classifications were shifted together to form the classification on the left. Since there were no values which appeared to represent trans males, they were not included in this data set.
        - Cis Female – Female, female, F, Woman, f, femake, woman, cis-female/femme, Female (cis), femail, Femake, Female
        - Cis Male – M, Male, male, m, maile, Mal, Male (CIS), make, Make, Man, msle, Mail, Malr, ostensibly male, unsure what that really means, Cis Man, cis male
        - Non-binary – Male-ish, something kinda male?, queer/she/they, non-binary, Enby, fluid, Neuter, Genderqueer, Androgyne, Agender, Guy (-ish) ^_^, male leaning androgynous, queer
        - Trans Female – Trans-female, Trans woman, Female (trans)

<u>Part II – Issues of Underrepresentation</u>

- Gender Representation
    - Of the 746 respondents, nearly 75% were identified to be cis men and roughly 25% identified as cis women. Only .011% of participants were identified as trans women or non-binary individuals. While these results suggest severe underrepresentation of female and non-binary individuals, the National Center for Women and Information Technology reports, only 26% of the computing workforce were women as of 2017. No reports exist outlining the percentage of

trans people who work in the technology sector. As of June 2016, however, the Williams Institute estimates that 0.6% of the American adult population is transgender. Thus, the miniscule representation of those who are trans in the survey likely representative of the general population of trans people in technology.

- State Representation
    - In total, respondents only came from 44 states and DC (excluding Alaska, Arkansas, Delaware, Hawaii, Montana and North Dakota). Considering that these states are some of the least populous ones, it makes sense that there are no respondents from these states.

Part III – Feature Deletion Based on Model Creation with Coworkers as a Target

- Using A 1R Model for Coworker Feature Deletion
    - The 1R model's accuracy was 67.23%. All 39 rules related to the state which determined the likelihood someone would tell their coworkers about their mental health. When tested against the training data, only 53.85% of instances were classified correctly.
- Using A Naïve Bayes Model for Coworker Feature Deletion
    - Overall, this model's accuracy was 57.19% and the features with highest probabilities included age, self-employment, working remotely, working at a tech company, and observed consequences. This model was equally as terrible as the first as it only predicted 41.81% of instances correctly.
- Using A RIPPER Model for Coworker Feature Deletion
    - In total, the RIPPER model classified 73.67% of instances correctly. The following rules were created as a result:
        - If (mental_health_consequence = No) and (mental_health_interview = Maybe) -> coworkers = Yes
        - If (Age >= 32) and (Age <= 33) and (phys_health_interview = Yes) -> coworkers = Yes
        - If (mental_health_consequence = Yes) and (no_employees = More than 1000) and (obs_consequence = No) -> coworkers = No
        - If (Age >= 43) and (treatment = No) and (Age <= 48) -> coworkers = No
        - Else coworkers = Some of them
- Using a C4.5 Model for Coworker Feature Deletion
    - The model classified 75.63% of instances correctly and created a tree with 21 leaves. Features used in the tree included mental_health_interview, work_interfere, anonymity, mental_health_consequence, care_options, benefits, remote_work and family_history.
- Using a Set of Random Forest Trees for Coworker Feature Deletion
    - The average out of bag error rate for the random forest models was 33.11% (yielding approximately 66.89% accuracy). Features which had a mean decrease Gini index greater than 10 were considered to be important: state (22.33),

self_employed (42.61), work_interfere (12.38), no_employees (17.9), leave (12.86), mental_health_consequence (18.62), and phys_health_interview (10.57).

- Feature Deletion
    - Since the following features were not considered important by any of the models, they were deleted from the coworker feature subset as follows:
        - Gender
        - wellness_program
        - seek_help
        - leave
        - phys_health_consequence
        - mental_vs_physical
    - The Naïve Bayes model had terrible accuracy. Since the following features were considered important in it (but not within the other models), they were also deleted:
        - tech_company

Part III – Feature Deletion Based on Model Creation with Supervisor as a Target

- Using A 1R Model for Supervisor Feature Deletion
    - The 1R model's accuracy was 54.9% and only had 3 rules based on mental_health_consequence as follows:
        - If (mental_health_consequence = Maybe) then supervisor = Some of them
        - If (mental_health_consequence = No) then supervisor = Yes
        - If (mental_health_consequence = Yes) then supervisor = No
- Using A Naïve Bayes Model for Supervisor Feature Deletion
    - Overall, this model's accuracy was 57.86% and the features with highest probabilities included age, gender, self_employed, remote_work, tech_company, anonymity, phys_health_consequence, mental_health_interview and obs_consequence.
- Using A RIPPER Model for Supervisor Feature Deletion
    - The RIPPER model's accuracy was 54.62% and only formed 4 rules as follows:
        - If (mental_health_consequence = Maybe) and (obs_consequence = Yes) -> supervisor = Some of them
        - If (anonymity = Don't know) and (wellness_program = Don't know) and (mental_health_consequence = Maybe) -> supervisor = Some of them
        - If (mental_health_consequence = Yes) -> supervisor = No
        - Else (supervisor = Yes)
- Using a C4.5 Model for Supervisor Feature Deletion
    - The C4.5 modeling algorithm yielded 61.9% accuracy with 73 leaves and used mental_health_consequence, obs_consequence, state, mental_health_interview, phys_health_interview, phys_health_consequence, work_interfere, family_history, wellness_program, mental_vs_physical, care_options, and treatment.

- Using a Set of Random Forest Trees for Supervisor Feature Deletion
    - This model yielded 58.17% accuracy. Features which had a GINI index of greater than 10 were considered important and listed as follows:
        - Age (26.51)
        - State (50.71)
        - work_interfere (14.04)
        - no_employees (21.54)
        - care_options (10.19)
        - seek_help (10.1)
        - leave (15.89)
        - mental_health_consequence (35.47)
        - phys_health_interview (11.49)
        - mental_vs_physical (14.47)
- Feature Deletion
    - Since the following features were not considered relevant by any of the models, they were deleted from the subset used to predict supervisor classification.
        - benefits