ML for Mental Health Awareness

Ankita Devasia & Lydia Yu

Significance & Objective

- Mental health disorders account for many of the most prevalent causes for disability in developed nations
- There is still widespread negative social stigma surrounding mental health
- 50% of individuals with severe mental health disorders go undiagnosed
- Many of these issues go unnoticed until they are very severe
- Providing people with a customized tool to predict their need for mental health treatment can bring greater awareness to and destigmatize these issues

Data

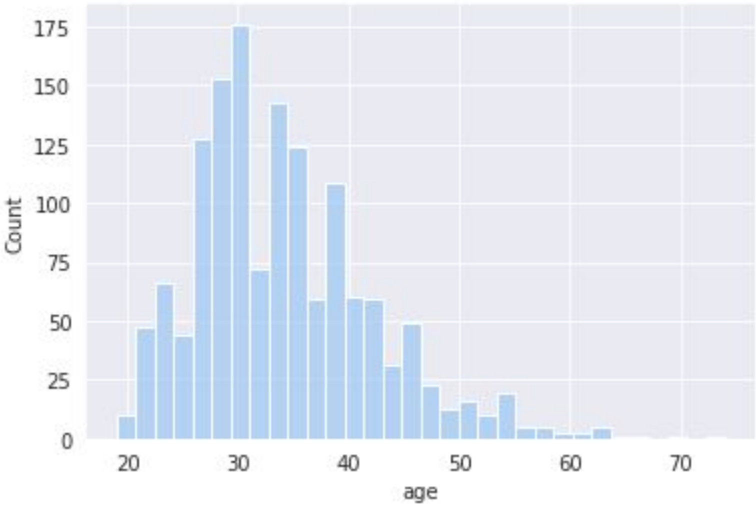
- Kaggle dataset from a 2016 survey conducted by the OSMH (Open Sourcing Mental Health)
- 1400+ responses from employees in tech on 63 questions related to their mental health and workplace environment
- Variables include:
  - Incidence of mental health disorder
  - Family history of mental health disorders
  - Work position
  - Ability to seek mental health resources at work

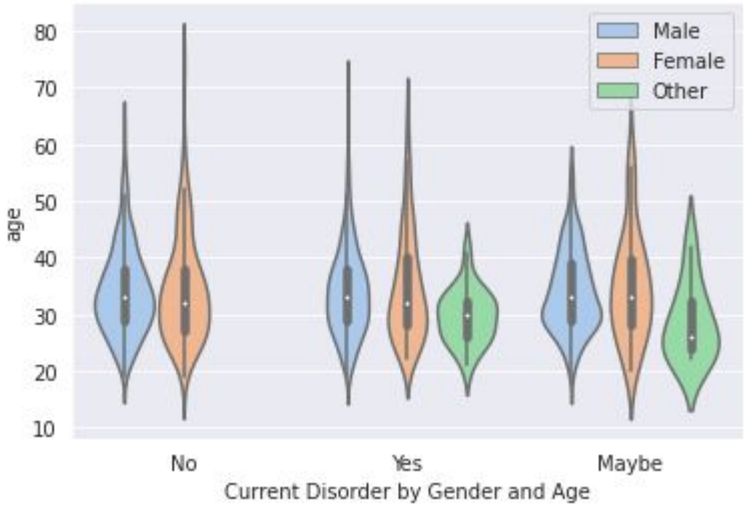
Data Preprocessing

1. Renaming columns from survey questions to shortened forms
2. Removing age outliers
3. Removing unimportant features
4. Imputing other NA values with the most frequent value
5. Encoding categorical variables
6. Encoding the dependent variable “mh_disorder_current”
    a. Yes/Maybe = 1, No = 0
7. Filling in missing values for “tech_flag” variable
8. Randomly split the resulting dataset into 70% train and 30% test

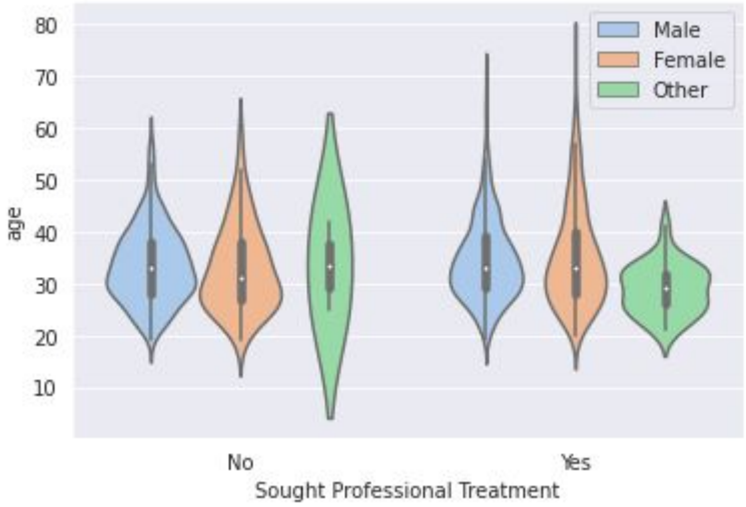
EDA: Demographic Information

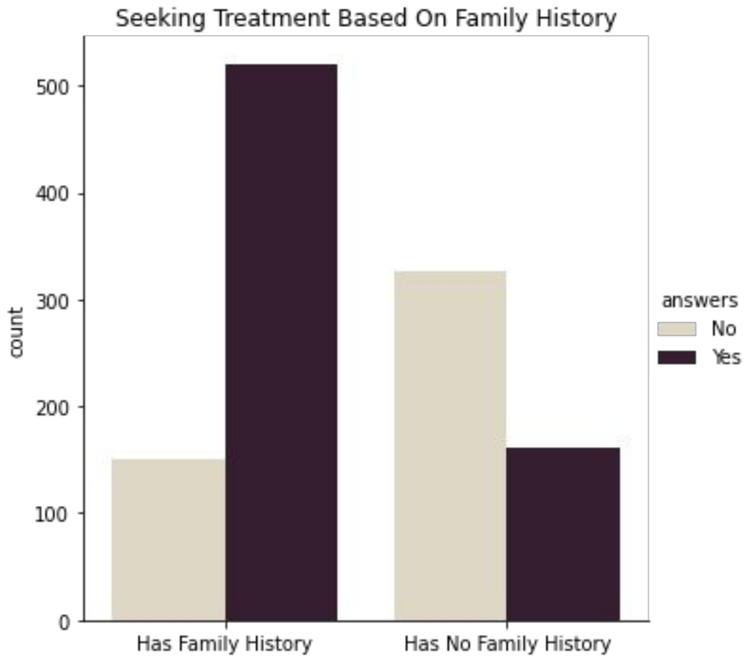
Gender	Count
Male	1060
Female	343
Other	27





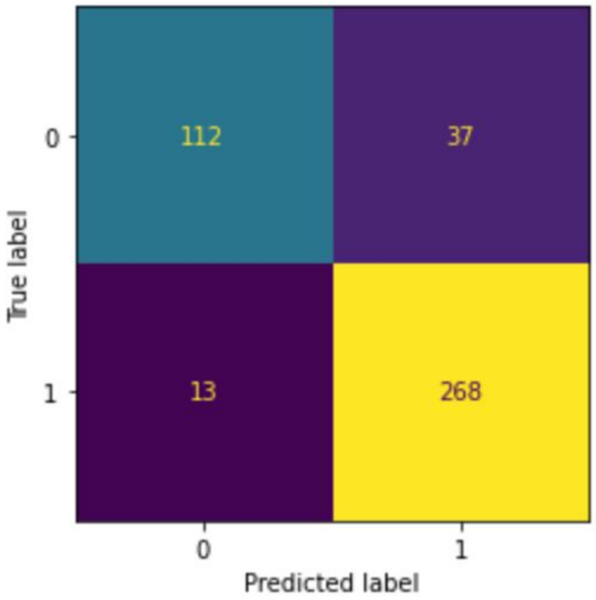
EDA: Seeking Professional Treatment



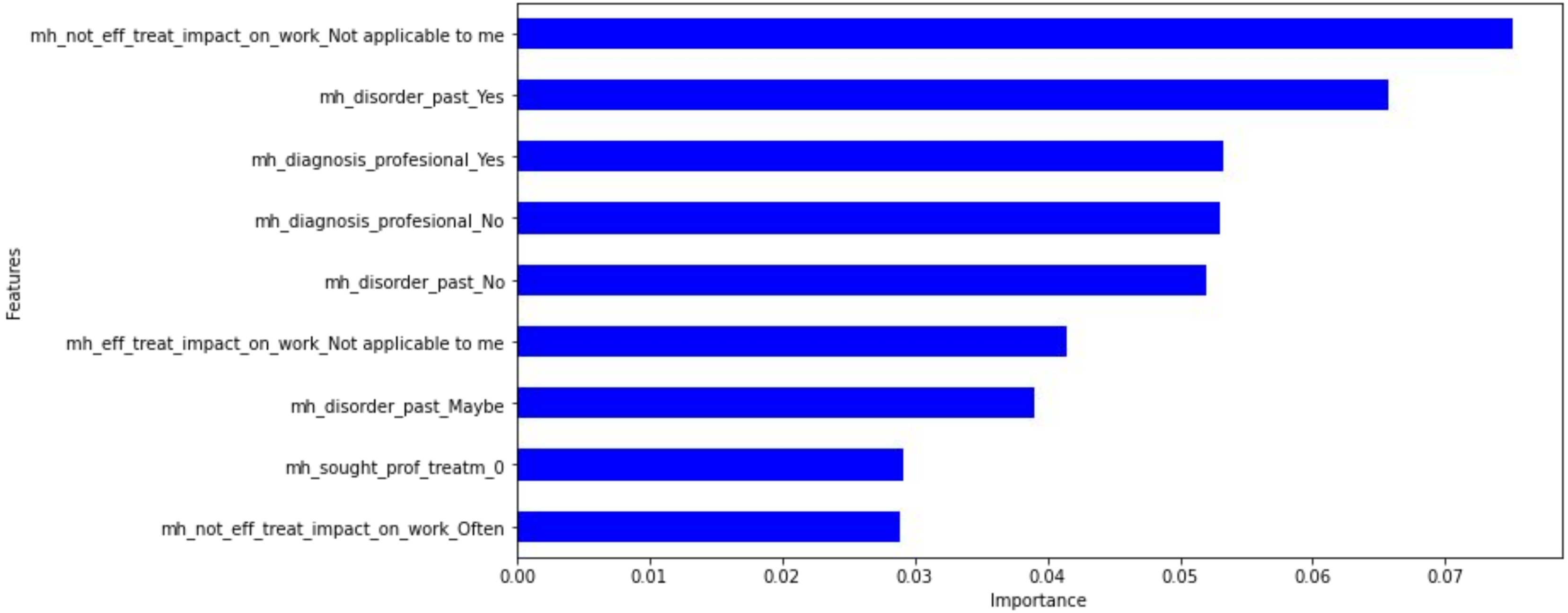


Random Forest

- In-sample R^2: 0.969
- Out-sample R^2: 0.884

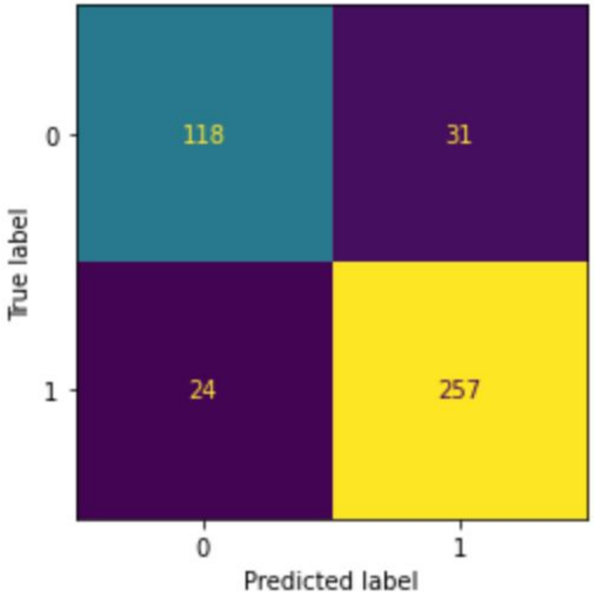


Random Forest Variable Importance



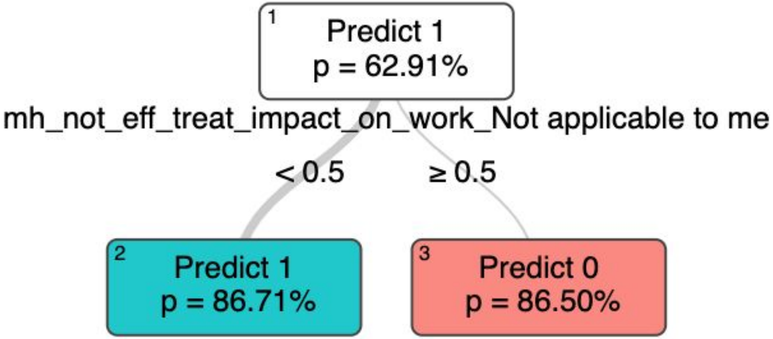
XGBoost

- In-sample R^2: 0.879
- Out-sample R^2: 0.872



Optimal Classification Tree

- Test AUC: 0.832



Comparing the Models

- RF had higher R^2 and lower false negatives than XGBoost
  - Variable importances show that the models used very obvious predictors
- OCT was very interpretable, not much less accurate, but too simple

Limitations and Further Work

- Output variable is whether or not they currently have a mental health disorder, not whether or not they should seek treatment.
  - Possible to need treatment even without having a disorder
- Survey data can be biased and have non-standardized scaling of responses
- No data on physical health
- Limited to tech company employees

Optimal Policy Trees

- Assign best treatment (seek professional treatment or not?) based on individual’s personal features
- Conduct follow-up survey on how much assigned treatment improved condition (outcome)
- Train OPT on these treatments and outcomes
- Can extend this concept to multiple different types of treatments to generate personalized prescriptions for different individuals
- Would require much more accurate data

Conclusion

- Survey data is not sufficient in predicting mental health disorders
- Results in models that give “obvious” results
- More objective and accurate data is needed, but difficult to obtain

References

- https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics
- https://mentalillnesspolicy.org/consequences/percentage-mentally-ill-untreated.html
- https://osmhhelp.org/research