Mental Health Classifier

/github.com/pschmeiser
in /in/pschmeiser
pschmeiser@mac.com

Background

Over the past 10 years mental health has finally gained more attention and importance in society.The occurrence of a mental health issues can not only cause major issues in one's personal life but can also disrupt the success of any company. As we are focused on data science this project chooses to look at those issues specifically in the tech industry. There are many factors that can cause the occurrence of a mental health issue which need exploration.This project uses the 2014 data set collected from 1260 people currently working in the tech industry. This is part of a multiple wave study that was continued in 2016. That data is not available at this time so it was not used for this case study.

Objectives

- *The first objective for this capstone was to explore the data. Each feature was explored and for some it was binarized.*
- *What factors lead to mental illness?*

Methods

A box plot on each of the features was examined to see if they had any influence on the the mental health classification. Almost all the features were influential so they all ended up in the model. They then had to have some processing before they could be used. Once processed two models were used for this data.

I tried two models: Random Forest and Gradient Boost. The results are in the next section.

Measures

As it is important to eliminate false negatives, the metric selected for this data set and model was recall. Recall focuses on eliminating our false negative rate. Additionally, precision is used as it keeps down the false positive rates.

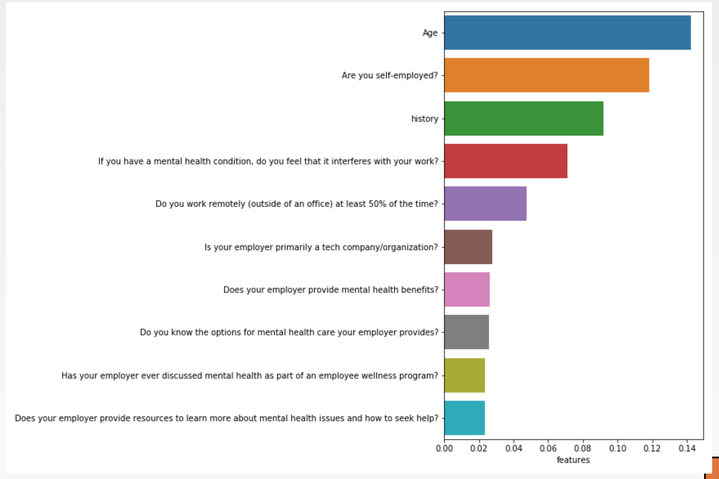
$$recall = \frac{true\ positives}{true\ positives\ + false\ negatives} \qquad precision = \frac{true\ positives}{true\ positives + false\ positives}$$

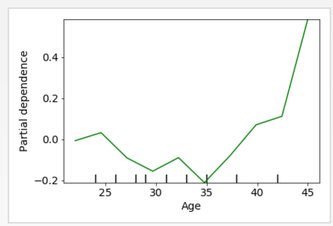
Results

Here are the scores from both the random forest model and Gradient Boosted

	Precision	Recall	F1
Random Forest	0.72	0.79	0.72
Gradient Boosted	0.74	0.79	0.76

Top 10 Feature Importances





Conclusion

As is shown by the feature importance the most influencing factor in determining potential for a mental health issue is age.

Following that we see that the most important feature is self-employment. Which surprisingly shows up before a history of mental illness

Obviously, a history of mental illness is going to be an important feature.

As we look at the feature importance a lot of influence is on communication. From the data it seems willingness to communicate sums up to a higher chance of mental illness from looking at the features.

References

1. *https://www.understandingsociety.ac.uk/documentation/health-assessment*