Mental Health Classifier

/github.com/pschmeiser
/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 in 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 persons currently working in the tech industry. The goal was to explore three different metrics and their relation to the occurrence of a mental health issue.

Objectives

- *The first objective for this capstone was to explore the data. Each feature was explored and for some it was digitized.*
- *The second objective was to then save the cleaned data in a new csv file*
- *To fit a random forest model*
- *To fit a Gradient Boosted model*
- *Use both models to output feature importance*
- *Use the gradient boosted model to create partial dependence plots*

Methods

Two methods were used to classify the mental health data.

The first used was the RandomForest Classifier. The results from the selected metrics will be show later.

The second used was Gradient boosted classifier.

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

```
accuracy: 0.740
precision: 0.725
recall: 0.787
============================================================
classification_report:
              precision    recall  f1-score   support

           0       0.76      0.69      0.72       205
           1       0.72      0.79      0.75       211

    accuracy                           0.74       416
   macro avg       0.74      0.74      0.74       416
weighted avg       0.74      0.74      0.74       416
```

Random Forest Model

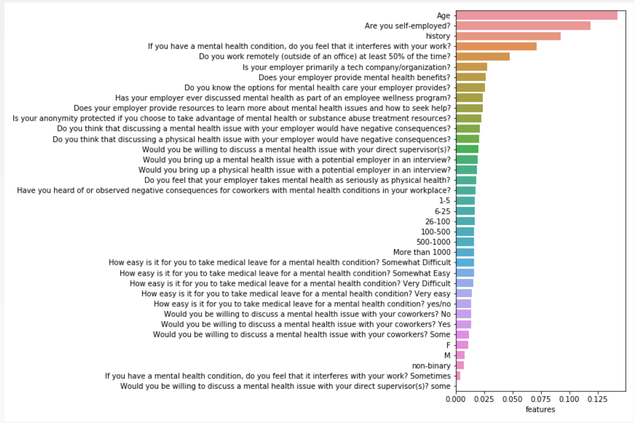
```
accuracy: 0.752
precision: 0.739
recall: 0.791
============================================================
classification_report:
              precision    recall  f1-score   support

           0       0.77      0.71      0.74       205
           1       0.74      0.79      0.76       211

    accuracy                           0.75       416
   macro avg       0.75      0.75      0.75       416
weighted avg       0.75      0.75      0.75       416
```

Gradient Boosted Model

Discussion



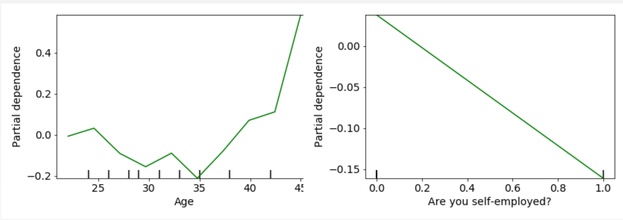
Conclusion

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