## Name

TheDebugger

## Question 1)

Training MSE: 0.8058427982143866

Training R^2: 0.42367291145486696

Model Used: Random Forest

Applied GridSearch to test different depths and number of estimators to prune the best possible random forest classifier. Ended up with depth of 12 and number of estimators = 100.

## Confusion Matrix:

A graph with numbers and a number

Description automatically generated

## Question 2)

Figure 1:

## A graph with red bars Description automatically generated

Figure 2:

A diagram of heat map

Description automatically generated

Figure 3:

A group of graphs with different sizes and colors

Description automatically generated with medium confidence

Figure 4:

A screenshot of a graph

Description automatically generated

Figure 5 (custom):

A diagram of different colored shapes

Description automatically generated with medium confidence

For this custom graph, I created violin plots showing the distribution of sentiment intensity scores for the cons section, pros section, and headlines section based on the overall rating given by the user in the training data. Some clear implications are the high positive intensity of pros section for high quality ratings, and the extreme low positive intensity of cons section for low ratings.

## Question 3)

## Actionable Insights:

For this question, I used my model to separate the most extreme results. I then extracted some of the most common phrases across these observations using a word map. This gave me a resultant list of common things people say in the cons section for very low ratings and pros section for very high ratings. I eliminated duplicates or similar ones and extracted the following common phrases or topics mentioned. I further investigated and investigated corresponding entries in the dataset to confirm my conclusions.

For High Rating reviews (4 and up), the following phrase map and counts table was generated for the pros text:

A white background with words

Description automatically generated A screenshot of a computer screen

Description automatically generated

For Low Rating reviews (2 and down), the following phrase map and counts table was generated for the cons text:

A white background with words

Description automatically generated A screenshot of a computer screen

Description automatically generated

Based on this data, we see that most of the low rating complaints (cons) are regarding mangers, work life balance, staff turnover rates, and long work hours. We also see that most of the high rating pros are regarding work life balance, work culture, flexible work hours, and great benefits. Thus, to improve overall ratings or to maintain high ratings, it is suggested that the firm tries to do the following:

1. Provide flexible hours and more benefits.
2. Nurture a collaborative and inclusive work culture/environment.
3. Train and hire effective managers that know how to lead.
4. Make attempts to reduce turnover rate or investigate why it happens more.

## Question 4)

ChatGPT helped me construct the phrase and word clouds to provide actionable insights. There are screenshots showing the questions I asked and the answers.

Question 1:

A black background with white text

Description automatically generated

Answer:

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

Follow-up:

A black background with white text

Description automatically generated

Answer:

A screenshot of a computer program

Description automatically generated