3-2 Assignment: Establishing a Baseline and Hypothesizing

Sarah Steinbaum

Southern New Hampshire University

DAT430: Leverage Data for Org Results

Dr. Matthew Keane

May 21, 2023

The HR department wants to measure how well employees are doing with the training and if further training would be beneficial to improve employee performance and job satisfaction. The questions I will address for this analysis are: Does training have an effect on job satisfaction and/or employee performance? If so, would further training assist in improving metrics for either of these two variables?

First, I opened the HR Training dataset and cleaned it to only provide me with data that is beneficial to this analysis as well as data that may be useful for further analysis. I imported pandas and then uploaded the cleaned data set and named it hr\_training. I then used the head() function to display the first 5 rows of data in this set as displayed below. JobSatisfaction and PerforamceRating are scored from 1 to 5 with 5 being a good rating and 1 being poor. The variable, training, displays how many trainings an employee received on a scale from 1 to 6.

The null hypothesis is that no relationship exists between training and job satisfaction or performance rating. The alternative hypothesis is that there is a collelation between training and at least one variable between job satisfaction and performance rating.

A screenshot of a computer

Description automatically generated with medium confidence

Next, I imported *matplotlib.pyplot* in order to create a scatter plot as this type of chart provides a clear visual that can portray the correlation between job satisfaction and training as well as performance rating and training. Matplotlib is a visualization library that is used to create charts such as the scatterplots displayed below.



I then created a scatterplot using the training variable on the y-axis and job satisfaction on the x-axis. This plot shows that there is a positive correlation between these two variables. The job satisfaction rating is lower when the amount of training the employee has had is low. Alternatively, if the employee had a high job satisfaction score, they had at least 4 trainings.



A picture containing screenshot, rectangle, line, diagram

Description automatically generated

I created another scatterplot using the training variable on the y-axis and performance rating on the x-axis. A positive correlation also exists between an employee’s performance rating and the amount of training. If an employee had a small amount of training, their performance was low. However, if the employee had more training, their performance rating would be higher.



A picture containing screenshot, rectangle, line, diagram

Description automatically generated

Once the scatter plots were created, I created a training model using the variables for job satisfaction and performance rating on the x-axis and the training variable on the y-axis.



Scikit-Learn is a machine learning library in python that provides “diverse algorithms for classification, regression, clustering, and dimensionality reduction” (Datagy, 2022). I used sklearn.model\_selection to import train\_test\_split. Using this code allows python to select the most appropriate model. Using train\_test\_split allows for the data to be trained and tested. The code in the second line labels the variables and formats the test size to be 30%.



The table for the x\_train is displayed below. This contains the variables for job satisfaction and performance rating.

A picture containing text, screenshot, number, font

Description automatically generated

The table for y\_train is displayed below. This contains the variable, training. To limit confusion, the variable, training, represents the number of trainings an employee had while y\_train displays the set of data that will be used in this machine learning process.

A screenshot of a computer

Description automatically generated with medium confidence

Next, I created a linear regression model using the sklearn.linear\_model. The values in both x\_train and y\_train are used for this linear regression model. The fitted values for this model are displayed below.

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing text, document, black and white, paper

Description automatically generated

The score() function allows python to display the accuracy of the linear regression model. This model is roughly 92% accurate. We have sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis. There is a correlation between training and both variables for job satisfaction and performance rating. More training would statistically improve employees’ job satisfaction and overall performance. A deeper analysis could tell us exactly how much training would be beneficial before the variables start to plateau.

A picture containing text, font, white, screenshot

Description automatically generated

References

*Introduction to scikit-learn (sklearn) in python • datagy*. datagy. (2022, January 5). https://datagy.io/python-scikit-learn-introduction/