Exploratory Data Analysis

Part 2

An exploratory data analysis, or EDA, is an approach used to investigate and understand the main characteristics, patterns, and relationships within a dataset. An EDA was conducted to understand what variables contribute to attrition within a growing organization. The objective of the EDA was to create a baseline using data provided by HR that management can use to determine areas they can improve needed to encourage employees to stay within the organization. The baseline data set for the exploratory data analysis was uploaded and the first five rows are displayed below using the head() function.

A screenshot of a computer

Description automatically generated with medium confidence

I imported a few libraries in addition to pandas on python to create two heatmaps in order to visually determine the correlation of the variables to *Attrition*. The first heatmap shows the variables with the highest positive correlation closer to red on the ROYGBIV scale and the highest negative correlation in dark purple. The second heatmap shows correlation on a gradient scale in blue with a darker shade indicating a strong positive correlation. Both heatmaps show that *DistanceFromHome* has the largest positive correlation and *PerformanceRating* has the largest negative correlation. Additionally, the variables *YearsSinceLastPromotiom*, *YearsAtCompany*, and *HourlyRate* have a decent effect on attrition. I removed several variables that did not show at least 0.13 or -0.13 on the correlation matrix and then created another gradient map in purple using these variables.

A picture containing text, font, screenshot, white

Description automatically generated



A picture containing text, screenshot, colorfulness, number

Description automatically generated

A picture containing text, font, screenshot, line

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

The following line graph demonstrates the effect of attrition on the variables *PerformanceRating* and *DistanceFromHome*. The blue line indicates no attrition, and the orange line does indicate attrition. The other line graphs portray the variables *WorkLifeBalance, DistanceFromHome, PerformanceRating*, and *YearsSinceLastPromotion* against Attrition individually. The graphs show that as *WorkLifeBalance* and *PerformanceRating* increase, attrition decreases. It also shows that as *DistanceFromHome* and *YearsSinceLastPromotion* increase, attrition also increases.



A picture containing text, plot, diagram, font

Description automatically generated

A picture containing text, line, diagram, plot

Description automatically generated A picture containing text, font, screenshot, plot

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generated A picture containing text, plot, diagram, font

Description automatically generated

Two predictive models were created to test the accuracy of the model including a logistic regression model and a random forest. The logistic regression model was created using Scikit-Learn which is a machine learning library in python that provides “diverse algorithms for classification, regression, clustering, and dimensionality reduction” (Datagy, 2022). I set up the logistic regression model with *Attrition* on the y-axis and the variables *DistanceFromHome, YearsSinceLastPromotion, YearsAtCompany, HourlyRate, JobSatisfaction, WorkLifeBalance, PerformanceRating* on the x-axis using a test size of 0.3. This model’s accuracy is 0.81208 which indicates that the logistic model using these variables is roughly 81% accurate. I then used a random forest to further determine the accuracy of the model using the same variables as the logistic regression model. The accuracy of this model is 86%.

**Logistic Regression Model**

A screenshot of a computer

Description automatically generated with medium confidence

**Random Forest**

A screenshot of a computer

Description automatically generated with medium confidence

In conclusion, leadership wanted to determine what factors lead to attrition to determine what areas they can improve to retain employees. The analysis determined that the variables *DistanceFromHome, PerformanceRating*, *YearsSinceLastPromotion, YearsAtCompany, HourlyRate, JobSatisfaction,* and *WorkLifeBalance* have an impact on attrition. It is possible the organization can offer remote or hybrid options to decrease the number of employees that are leaving due to the distance and potentially improve their work-life balance. Additionally, adjusting the employee performance rating criteria and providing deserving employees with a good rating that includes an hourly raise and working toward a promotion should lead to increased job satisfaction. Since the random forest model had a higher accuracy percentage of 86% and would be less likely to have overfitting, I would recommend using this model over the logistic regression model.

References

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

Reddy, E. P. K. (2021, March 25). *Step by step process of feature engineering for machine learning algorithms in Data Science*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/03/step-by-step-process-of-feature-engineering-for-machine-learning-algorithms-in-data-science/

Sun, E. D., Ma, R., & Zou, J. (2022, December 30). *Dynamic visualization of high-dimensional data*. Nature News. https://www.nature.com/articles/s43588-022-00380-4

*Choosing colormaps in matplotlib#*. Choosing Colormaps in Matplotlib - Matplotlib 3.7.1 documentation. (n.d.). https://matplotlib.org/stable/tutorials/colors/colormaps.html