Statistical data analysis

IBM HR Analytics Employee Attrition & Performance
A general assignment for data scientists

Fall 2022

Introduction

You are provided with an artificial dataset on employee attrition assembeled by IBM data scientists. It comprises of 1470 employee records with 34 attributes, rangeing from former education, parameters of their job, to slightly touching some personal details. In the beginning, you are supposed to visualize the main relationships and trends. Your main task is to learn from the dataset and to understand the factors driving the attrition. Also, you should construct, test and evaluate a predictive model that tells whether or not an employee will quit.

The task can be found at: Kaggle. There are already some solutions to the task, mainly from the exploratory analysis perspective. Get inspired, as this aspect played only a supplementary role throught the SAN course. In further efforts, you are supposed to provide your own models and statistical tests you learned within the course.

1 Tasks

The main aim of this assignment is to demonstrate your ability to apply the methods that you learned during the course. The assignment can be decomposed into the following well-defined tasks:

1. **Exploratory analysis of the dataset**. Using visualizations, try to get as many insights from the data as you can. Preprocess the data if needed, carry out dimensionality reduction and/or clustering. The main goal is to explore the relationships between attrition and the remaining parameters. In dimensionality reduction and clustering, try to also interpret the new space and potential clusters.

2. **Attrition predictive model.** Create a model that predicts the attrition probability based on the remaining variables. Evaluate and compare the performance of the models, employ cross-validation to get an unbiased estimate.

If needed, utilize feature selection/engineering to simplify the models and improve their performance. Use at least one method that we touched in the course, but you can compare with other methods too (neural networks, gradient boosting trees, etc.). Discuss your choice of performance metrics and try to interpret what factors your models base their decision on.

3. **Creative section.** This section serves as a playground for you to try out other techniques you have learned during this course. Feel free to apply any model you find suitable (except for the ones used in the previous step). Extract the most useful conclusions and discuss your findings. Care for the correct usage of the models, and verify their assumptions if necessary. Methods you can use are (but not limited to): linear regression, GLMs, (M)ANOVA. Be creative!

4. **Discussion of the results.** Provide a verbal summary of your results. Compare with observations and conclusions from parts 1 and 2 (e.g. do features important in dimensionality reduction match those important for class separation?). Optionally, propose future work or potentially interesting tasks you could not solve with the given dataset.

2 Submission and evaluation

Submit your solution to the upload system. Place your RMarkdown report in the folder named *DSTask* to designate you have chosen the Data Science assignment. Make sure your reported can be knitted.

You can obtain up to 15 points for this assignment, 4 points each for the first subtasks plus 3 points for the final discussion. In the first three subtasks, approximately 70% will be given for the concept of the solution (selection of statistical methods and their correct application, depth of the solution), 20% for the answers that summarize your solution in the individual subtasks (interpretation of your results and explanation of their practical impact in natural language), 10% for formal issues (clarity of the code, comments, readability of the rmarkdown as a whole).