**Analysis Project**

**Addressing Employee Retention**

**(Due June 29th, 2024)**

# Introduction

For the class analysis project, we will be working with a data set from the field of Human Resources Analytics. Ideally, companies would like to keep attrition rates (the proportion of employees leaving a company for other opportunities) as low as possible as a means to build organizational capability and to reduce costs and business disruptions that come with replacing productive employees on short notice.

please submit the following for the analysis project:

* R Markdown file in based on the template
* Word document which has knit from the R Markdown file.

The data – “Employee\_Data” - consists of 1,100 employee records for a U.S. based product company. The rows in this data frame represent an employee’s attributes across the variables listed in the table below.

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Data Type |
| left\_company | Did the employee leave the company? (Yes/No) | Factor |
| department | Department within the company | Factor |
| job\_level salary weekly\_hours | Job Level (Associate - Vice  President)  Employee yearly salary (US Dollars)  Self-reported average weekly hours spent on the job (company survey) | Factor  Numeric  Numeric |
| business\_travel | Level of required business travel | Factor |
| yrs\_at\_company | Tenure at the company (years) | Numeric |
| yrs\_since\_promotion | Years since last promotion | Numeric |
| previous\_companies  job\_satisfaction performance\_rating | Number of previous companies | Numeric |
| for which the employee has  worked  Self-reported job satisfaction Factor (company survey)  Most recent annual performance Factor rating | |

|  |  |  |
| --- | --- | --- |
| Variable | Definition | Data Type |
| marital\_status miles\_from\_home | Marital status (Single, Married, or Divorced)  Distance from employee address to office location | Factor  Numeric |

**Project Requirements**

The project requirements are to:

Conduct exploratory data analysis on the data set using commands including dim( ), str( ), head( ), glimpse( ), skim( ) and summary( ).

Decide on three questions that explore the relationship between whether someone left the company and other variables in the data set. Using summary tables and graphs explore each of the three questions – for example using using dplyr/tidyr or ggplot or both and draw a conclusion as to whether a variable is related to whether someone leaves the company.

Fit two classification algorithms (logistic regression and one other) to predict the response variable of left\_company. You should use all of the other variables in the Employee\_Data dataset as predictor variables for each model. For the analysis, please follow the template in the Predictive Modeling section below.

Write a project report.

Knit the report with the code and summary tables and plots and turn in as both Word and Rmd files.

# Overview

Executives at this company have hired you as a data science consultant to identify the factors that lead to employees leaving their company. They would like for you to do an initial data analysis that is able to provide the insights on the likelihood of an employee leaving and to help address employee turnover.

Specifically, the broad question that the executive team is trying to answer is:

• What are the factors that contribute to employees leaving the company based on summary statistics?

# Data Analysis [30 Points]

In this section, please think of at least three relevant questions that explore the relationship between left\_company and the other variables in the Employee\_Data data set. The goal of your analysis should be discovering which variables drive the differences between who leave the company and those that do not.

Please each question and provide supporting data summaries with either a summary data frame (using dplyr/tidyr) or a plot (using ggplot) or both.

# Predictive Modeling [70 Points]

In this section of the project, you will fit two classification algorithms to predict the response left\_company. You should use all of the other variables in the Employee\_Data data set as predictor variables for each model.

Please follow the machine learning steps below.

The data splitting and feature engineering steps should only be done once so that your models are using the same data and feature engineering steps for training.

* Split the Employee\_Data into a training and test set (remember to set the seed)

* Specify a feature engineering pipeline with the recipes package

- You can include steps such as skewness transformation, dummy variable encoding or any other steps you find appropriate

• Specify a parsnip model object

- Choose from the following classification algorithms:

 Logistic Regression

 LDA

 QDA

 KNN

 Decision Tree  Random Forest.

* Package your recipe and model into a workflow

* Fit your workflow to the training data

- If your model has hyperparameters:

 Split the training data into 5 folds for 5-fold cross validation using vfold\_cv (remember to set your seed)

 Perform hyperparamter tuning with a random grid search using the grid\_random() function for Random Forest or regular tune\_grid for KNN

 Hyperparameter tuning can take a significant amount of computing time. Be careful not to set the size argument of grid\_random() too large. I recommend size = 10 or smaller.

 Select the best model with select\_best() and finalize your workflow

• Evaluate model performance on the test set by plotting an ROC curve using autoplot() and calculating the area under the ROC curve on your test data

# Summary of Results [50 Points]

Write a summary of your overall findings and recommendations to the executives at the company. Think of this section as your closing remarks of a presentation, where you summarize your key findings, model performance, and make recommendations.

Please include:

1. An introduction where you explain the business problem and goals of your data analysis

* What problem(s) is this company trying to solve? Why are they important to their future success?
* What was the goal of your analysis? What questions were you trying to answer and why do they matter?

2. Highlights and key findings from your Exploratory Data Analysis section

* What were the interesting findings from your analysis and why are they important for the business?
* This section is meant to establish the need for your recommendations in the following section

3. Your “best” classification model and an analysis of its performance

• In this section you should talk about the expected error of your model on future data

– To estimate future performance, you can use your model performance results on the test data **Report and Summary of Results**

Please follow the report template (please see below) with Introduction, Data and Model, Discussion and Analysis, Conclusions and Appendix sections.

**Data Analysis Project Template**

Name

Date

# Structure for Data Analysis Project Report

The data analysis report is specifically for the client, and should address the challenges faced by the client. You should not look at this report as a term paper in a class or a report about a subject. This document is meant, primarily, to start an organized conversation between you and your client. The document should have the following structure.

## 1. Introduction. The introduction should include

* Describe the problem. What substantive question are you trying to address?
* Summary of the study and data, as well as any relevant substantive context, background, or framing issues.
* The “big questions” answered by your data analyses, and summaries of your conclusions about these questions.
* Brief outline of remainder of paper.

## 2. Data and Model

Divide the section into the following categories.

* Data
* Methods
* Analysis
* Results

For every analysis, start with a question you want to address, and provide a detailed response using data, methods used, and finally the results.

**3. Discussion and Analysis**

## 4. Conclusions

The conclusion should reprise the questions and conclusions of the introduction, augmented by some additional observations or details gleaned from the analysis section. New questions, future work, etc., can also be provided in this section.

**5. Appendix/Appendices.**

This is the section, where you will provide the detailed information on the following.

* Detailed tables from the R output
* Plots from R output
* R commands