STA545: Project Proposal - Analysis of HR Employee Attrition and Performance Dataset

Arun Krishnamurthy (50247445) (Class No: 34)
& Ashwin Vijayakumar (50249042) (Class No: 03)

November 1, 2017

1 Abstract

The idea of the project is to perform exploratory data analysis on the HR Attrition and Performance sample data set given IBM's Watson Analytics. This helps us say if an employee is going to quit the company given certain attributes, and as a result provide some insight about retaining valuable employees.

2 Dataset

This dataset is a fictional sample dataset created by IBM scientists at Watson Analytics. It can be downloaded at: https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/. It has 35 different attributes on 1470 employees. These are the attributes:

Age	Attrition	BusinessTravel
Department	DistanceFromHome	Education
EmployeeCount	EmployeeNumber	EnvironmentSatisfaction
HourlyRate	JobInvolvement	JobLevel
JobSatisfaction	MaritalStatus	MonthlyIncome
NumCompaniesWorked	Over18	OverTime
PerformanceRating	RelationshipSatisfaction	StandardHours
TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance
YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
DailyRate	EducationField	Gender
JobRole	MonthlyRate	PercentSalaryHike
StockOptionLevel	YearsAtCompany	

3 Approach

3.1 Pre-processing

Here, we split the data into train and test sets, and perform any other pre-processing techniques, such as outlier detection, etc., that may be necessary to bring the dataset to a point where data mining tasks can be performed on it.

3.2 Exploring the Data

In this step, we will build various plots for the data to glean any obvious, inherent relationship that may exist in the data. This will inform our decision making in the next steps when we build the model to make the required predictions.

3.3 Building and Validating the Model

This is a multivariate linear regression problem with output variables as Attrition and PerformanceRating. Our aim is to fit a regression model with cross-validation to predict: whether or not the employee is going to quit, and, the performance rating of the employee. We also employ various other statistical tools such as model selection techniques (shrinkage, dimensionality reduction, and subset selection), and resampling methods (cross validation and bootstrap methods) to enhance our model.

4 Presentation and Visualization

We present our analysis of the data and model created, using various plots and tables to help visualize the insights and results obtained from the data. Through this process, we will be able to successfully predict, to a reasonable accuracy, whether an employee may or may not quit the company, along with their performance rating which will help the company retain valuable employees.