Predicting Attrition of IBM Employees

Prepared for: DATS6103 Introduction of Data Mining Final Project
Prepared by: Group 1, Reema Domadia, Dinesh Kumar Padmanabhan, Utkarsh Nigam

Introduction:

The year 2015 welcomed a fifth generation of attributes to create the 21st century workforce. The youngest generations, millennials and generation-Z, came with an attribute that dramatically revolutionized the workforce: the computer age. The introduction of remote work enabled employees to find jobs that they enjoyed and fit their lifestyle, transforming the way companies conducted business. Within a decade, employers saw a reverse effect in which companies are now having to appeal to their employees. While the workforce saw an increase in the candidate pool, accordingly, they also saw an increase in employee departures creating a movement of what is now called 'job hopping'. According to a study called "*Should I stay or should I go*?", IBM's Dr. Sheri Feinzig and Dr. Haiyan Zhang compared driving values between Millennials, Generation X and the Baby Boomers when jumping positions and found that while 'better compensation and benefits' was a primary driver, valued between 70-78%, the second leading influencer for millennials was 'better career development opportunities' at 77% while Generation X and baby boomers looked for' job security' at around 70% (Dr. Sheri Feinzig and Dr. Haiyan Zhang, 2016). Such insights are crucial to human resources departments as they are constantly looking for ways to decrease attrition of the employees in their companies.

Problem Statement and Dataset:

The purpose of this research project is to identify contributing factors to enable retainment and discourage employees from 'job hop'. Due to the complexity of this project, it is not feasible to analyze the entire workforce. Consequently, through a dataset we found on Kaggle.com, we decided to study a company constantly updating its policies and talent acquisition strategies to keep up with the changing needs of our time: International Business Machine (IBM). The dataset, as it was found, does not require any cleaning and is equipped to fuel the analysis of this project. The dataset consists of 35 variables and 1470 observations assessing the employee's attributes ranging from age and marital status to training time and time in position. Due to the abundance of data recorded from various aspects of an individual's life, this dataset enables us to deep dive using clustering techniques such as correlation to identify values that the employees find most attractive and characteristics that drive employees away. Through a quick exploratory data analysis (eda) for this dataset, we found the following statistics: 65% of employees in this dataset work in research and development and 30% in sales, 71% of employees rarely travel, 73% of employees have a background in life sciences or the medical field, 60% of employees are satisfied with their work environment, 60% of employees identify as male and 40% as female, 46% of employees are married and 32% are single. Through these statistics we can better understand the characteristics of our employee population.

Methodology:

To promote collaboration in analyzing this problem, GitHub will be used to compile all efforts made by members of the team. Additionally, we will utilize a combination of the PyCharm IDE and the Anaconda libraries, often referred to as Conda, to employ a combination of feature selection methods and predictive machine learning techniques. Due to the versatility of PyCharm, with the assistance of Conda libraries, the sole software is expected to suffice the technical needs of this project. Respectively, correlation enables the analysis of the strength relationship between the variables in the dataset while logistic regression, random forests, and decision trees will manufacture a model that will output attrition rates. These supervised machine learning algorithms were identified

as crucial algorithms for this project due to their strengths in training models in influencing decisions to output the respective binary results. Through this model, we will be able to identify the factors that play a key role in the companies' attrition rates and identify what employees are prone to abandoning their positions. This information is vital in the success of a company as the employees are the backbone of the company and their progression will lead to the success of the company. Additionally, this information will enable management to improve weaknesses in talent retention strategies in order to retain its employees.

Schedule:

To ensure our project is on track we will pursue with below schedule (Figure 1).

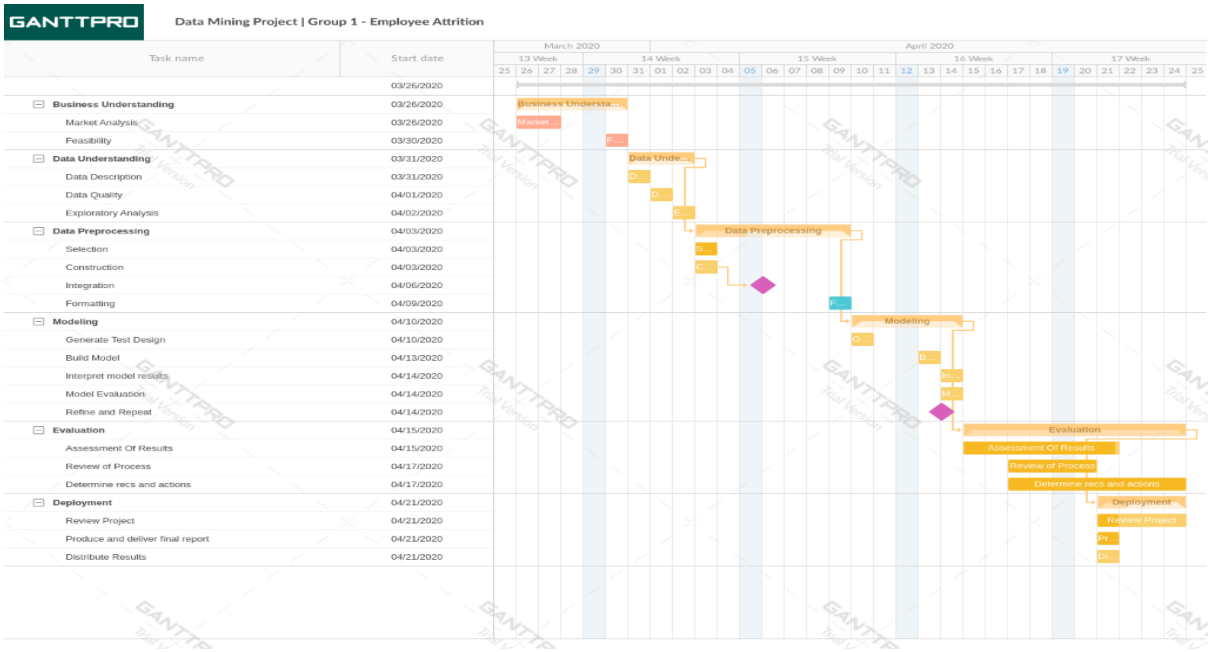

Figure 1: Schedule for Project

Evaluating Performance:

Through modelling in the PyCharm IDE, the resulting GUI will identify the dissatisfaction reasons for IBM employees based on the training variable. By checking model performance using a confusion matrix, we will be able to identify opportunities where it can be confused when making predictions. Recognizing errors can not only mitigate potential biases but also identify the associated type I and type II errors. By doing so, we will enable the model to output optimal results that will enhance leadership's ability to make decisions on mitigating increasing attrition rates. This coupled with pre-existing extensive research on this topic by IBM, we will be able to assess the validity of our model. The results of this study are outlined in a whitepaper titled *"Should I stay or Should I go?"* and a PowerPoint deck titled *"Predictive Retention: How to Know Before They Go"* by Haiyan Zhang, Ph.D. and Sheri Feinzig, Ph.D. Due to the level of analysis outlined in these documents, we believe it will provide the context we need coupled with our own industry experience and will help bridge our model output to real-world applications. In other words, while the observations in the study conducted by IBM may assess a different sample population, the insights and conclusions should hold true for the output of our model as well.

References

History.com Editors. “American Women in World War II.” *History.com*, A&E Television Networks, 5 Mar. 2010, www.history.com/topics/world-war-ii/american-women-in-world-war-ii-1.

Haiyan Zhang, Ph.D. and Sheri Feinzig, Ph.D. “Should I stay, or Should I go?” *IBM Smarter Workforce Institute.* IBM. https://www.ibm.com/downloads/cas/08GZQKL1.

Haiyan Zhang, Ph.D. and Sheri Feinzig, Ph.D. “Predictive Retention: How to Know Before They Go” *IBM Smarter Workforce Institute.* IBM. https://talentguard.com/wp-content/uploads/2017/10/TalentGuard-IBM-predictive-retention-2017.pdf