What Factors Inspires IBM Employees to Resign?

Reema Domadia
Dinesh Padmanabhan
Utkarsh Nigam
DATS6103
Professor Amir Jafari
April 28, 2020

Table of Contents

Introduction

According to an 2016 article by Alvernia University, the cost to replace an employee is as follows: entry-level, 30-50% of their annual salary; mid-level, 150% of their annual salary; and experts, cost up to 400% of their annual salary. Employees are the backbone of an organization and it is often observed that the individual employees success at the company equates to the organizations, as a whole, success. Employees, who are happy with their leadership, often work harder and are more dedicated to their work. As most employees are ambassadors for the company to their clients and customers, this often results in a trust between the employee, the firm, and the client. On the contrary, companies that struggle to retain their talent often see an impact not only in their revenue but also their company culture as well. When an employee leaves, other employees stop to ask why and in other cases, high turnovers, decrease work morale negatively impacting productivity. Additionally, in an age, where it is almost expected for an individual to jump positions for the first 10 years of their career, it can cost employers a fortune to hire and train new employees not to mention direct costs of advertising and interviewing candidates.

The job of attracting new talent comes with a parallel task of retaining talent. In a world where technology has connected the entire world, the options for companies and candidates are endless. The only true costs are time to fill positions and revenue lost due to vacant positions. Human resource departments across the world spend time and resources trying to attract talent that shares the company values and will contribute to the company's intellectual capacity. Fundamentally, there are two types of employees: firstly, those that are constantly looking for a challenge and opportunities to grow their skills and the second, those that are just trying to earn a living and/or support their families. Companies have adopted a number of attraction and retainment strategies such as promotions, merit salary raises, flexible and telecommuting work schedules, and training and education opportunities, to name just a few, to appeal to their employees. This project focuses on identifying variables that encourage employees to stay with the company and conditions under which attrition is experienced.

The following sections of this report will illustrate our dataset, methodology, analysis, and the results. The dataset segment will explain the source of our dataset and explain exploratory data analysis finding that shaped the course of our project. The methodology highlights the model's we've created and softwares we used to create the platform for our analysis. Finally, the analysis and results will showcase our findings and concluding remarks.

Dataset

Due to the complexity of the study, it is important to narrow down our scope by identifying a sample size. A larger population may not fortify the granular analysis we are hoping to derive. We understand that each company hires its employees based on how well the candidate will be able to fit into the company's unique culture, and that narrowing down the population to a specific sample size may not be representative of the entire workforce at large. Keeping this in mind, due to the time constraints of this project, the team firmly believes that the arguments for a larger dataset is outweighed by the arguments of a smaller sample size dataset. The team will process with a sample side dataset as we believe it is necessary to limit the scope of the project to ensure compliance with a specific, measurable, achievable, relatable, and time-oriented question. We hope that our analysis from this project will lay the foundation for a future, more elaborate study.

To initiate this study, we used an IBM attrition dataset found on kaggle.com to study factors and conditions that encourage employees to resign from their positions. As found, the dataset did not require any cleaning and was well equipped to serve as the backbone of this study. Originally, the dataset contained 1470 observations assessing 35 variables that ranged from demographics, work-life balance, employment history, investments, and job satisfaction. The following algorithms were used to derive the aforementioned insights:

```
attrition = pd.read_csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
print(attrition.head())
print(attrition.shape())
print(attrition.dtypes())
print(attrition.shape())
print(attrition.isnull().sum())
```

After preliminary analysis, we removed three variables that did not impact the analysis at large: 'Employee Count', 'Employee Number over 18', and 'Standard Hours'. The variables 'Standard Hours' and 'Employee Number of 18' are standard throughout the dataset with 'Employee count' acting as the index, thus just creating unnecessary noise in the dataset. The resulting dataset contained 31 variables with 1470 observations. By eliminating variables that cause noise in the dataset we are able to eliminate clutter that may swade the data and decrease the time required to analyse the dataset.

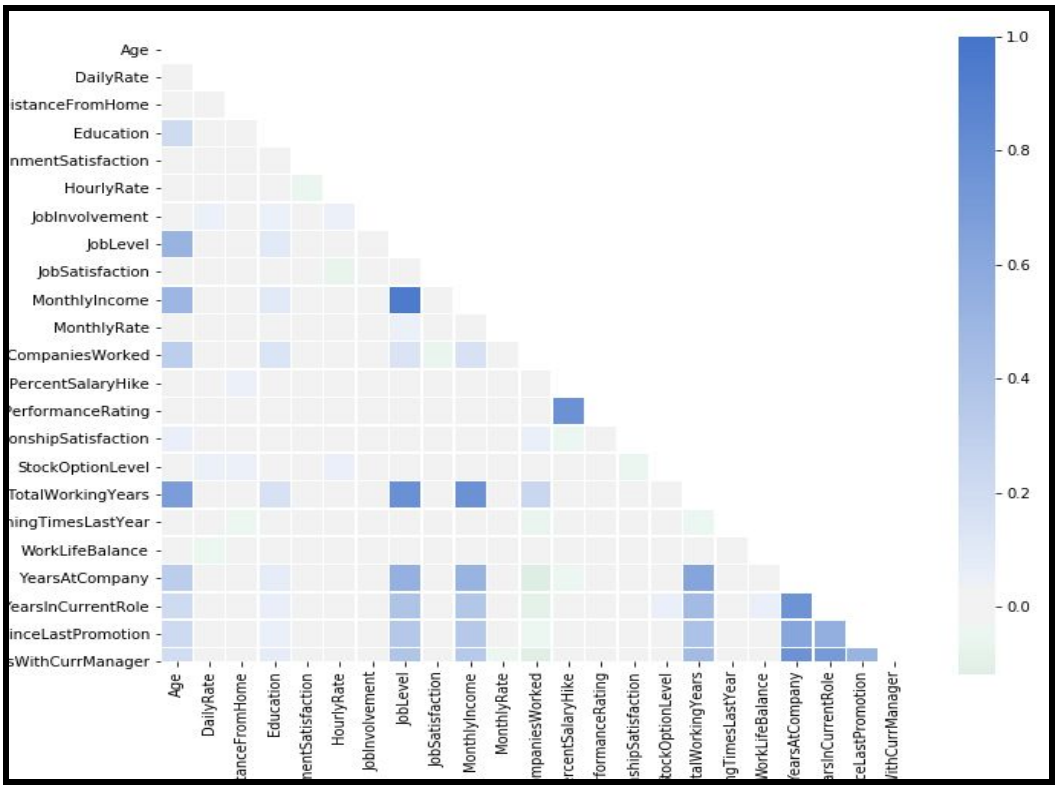


Figure 1: Correlation Plot

To set the framework for our exploratory data analysis (EDA) we used a correlation plot to gain a better high-level understanding of the relationships between both the ordinal and continuous variables. Correlation plots allow the assessment of the relationship between two variables. Through the correlation plot, we learned the following (Fig 1):

- ‘Monthly income’ is dependent on the ‘Job-Level’
- ‘Total working years’ impacts ‘Job Level’ and ‘Monthly Income’
- Salary Increases are dependent on employee performance
- ‘Age’ is a predictor of ‘Education’, ‘Job Level’, ‘Monthly Income’ and ‘Total Working Years’

Additionally, the following source code was utilized to generate Figure 1:

```
att = attrition.select_dtypes(include=[np.number])
corr = att.corr()
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
```

```
f, ax = plt.subplots(figsize=(11, 9))
cmap = sns.diverging_palette(-220, 255, as_cmap=True)
heatmap = sns.heatmap(corr, mask=mask, cmap=cmap, center=0.0,
                      vmax=1, square=True, linewidths=.5, ax=ax)
plt.savefig('corr-heat.png')
plt.show()
```

The abundance of data recorded from various aspects of an individual's life enables the identification of company values that the employees find most attractive and characteristics that drive employees away. Through a EDA analysis, we learned that the company is doing a good job retaining its employees, as they only saw an attrition rate of about 15% compared to the 1250 employees (out of 1470) that still work there (Fig 2).

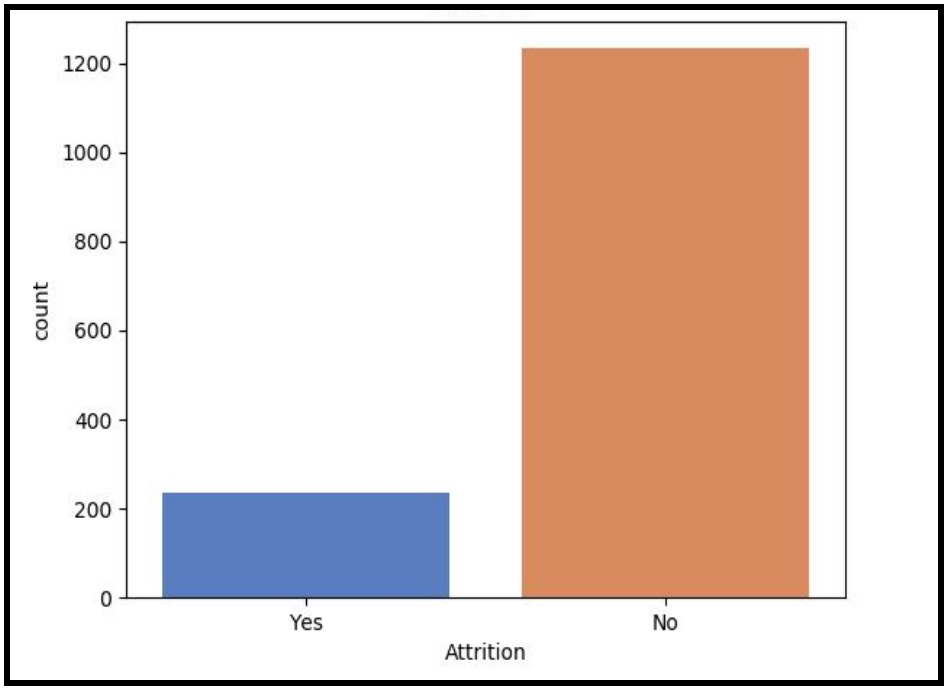


***Figure 2**: Attrition*

Analyzing the 'years at company' variable, we learned the company has fairly high turnover as about half the employees represented in this dataset have been at the company between 0-10 years and about a quarter between 10-20 years with the remaining distributed between 20-40 years (Fig 3). Additionally, we learned that the workforce at this company is fairly young with over half having worked a total of 12

years or less and only worked for one company (Fig 3). Through these high-level statistics we were able to set the framework for the rest of our analysis. We are interested in identifying what aspects and benefits of the company employees enjoy the most and areas where the company could do better to retain more employees.

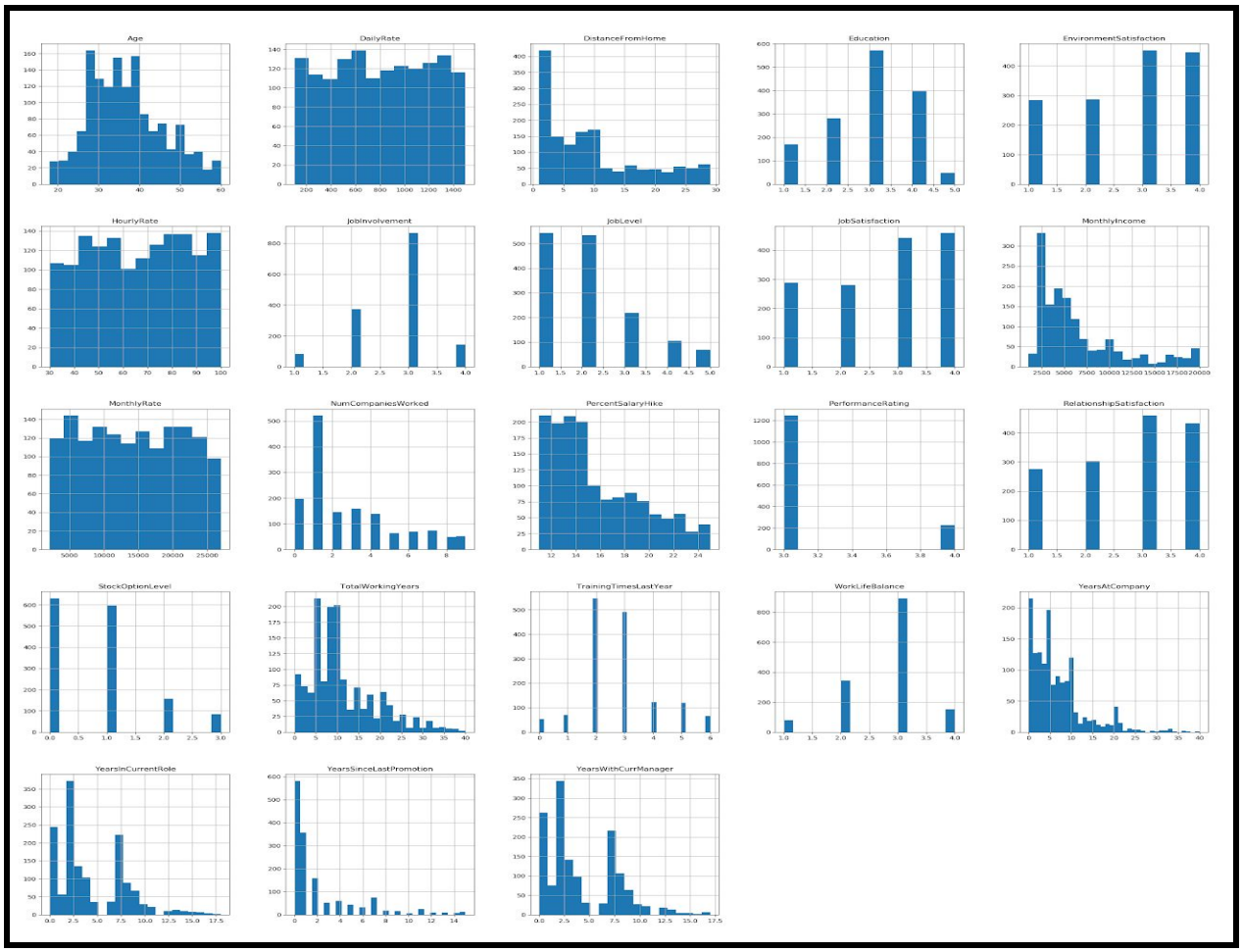


Figure 3: *Exploration Data*

Upon understanding the type and distribution of data recorded in our dataset, we were better able to derive insights that allowed us to set up a more robust framework. Through further analysis, we learned that the branch of IBM whose employees we are studying specializes on research and development with other departments such as sales and human resources as supporting departments (Fig 4). Upon learning this, we

believe either our conclusion is correct in that this is a satellite branch focusing on R&D or we are working with a partial dataset as we dont see other departments such as IT.

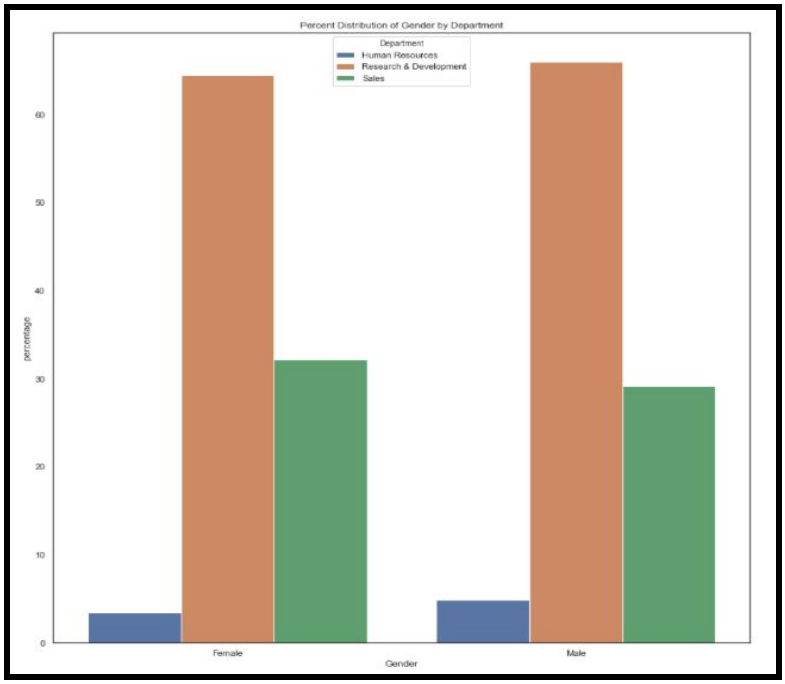


Figure 4: Gender Breakout of Department

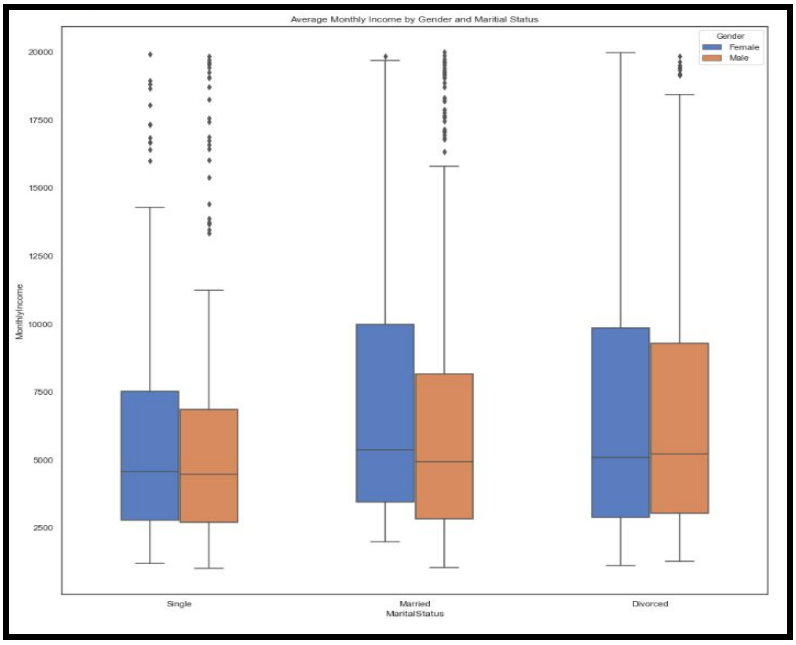


Figure 5: Marital Status, Gender Breakout Assessing Income

Furthermore, we are interested in understanding that income disparity between two demographic groups: gender and marital status. Through our analysis we learned that there is no discrimination between the genders or maritial statuses. We found that all groups average at about 50,000 USD for their monthly income (Fig 5).

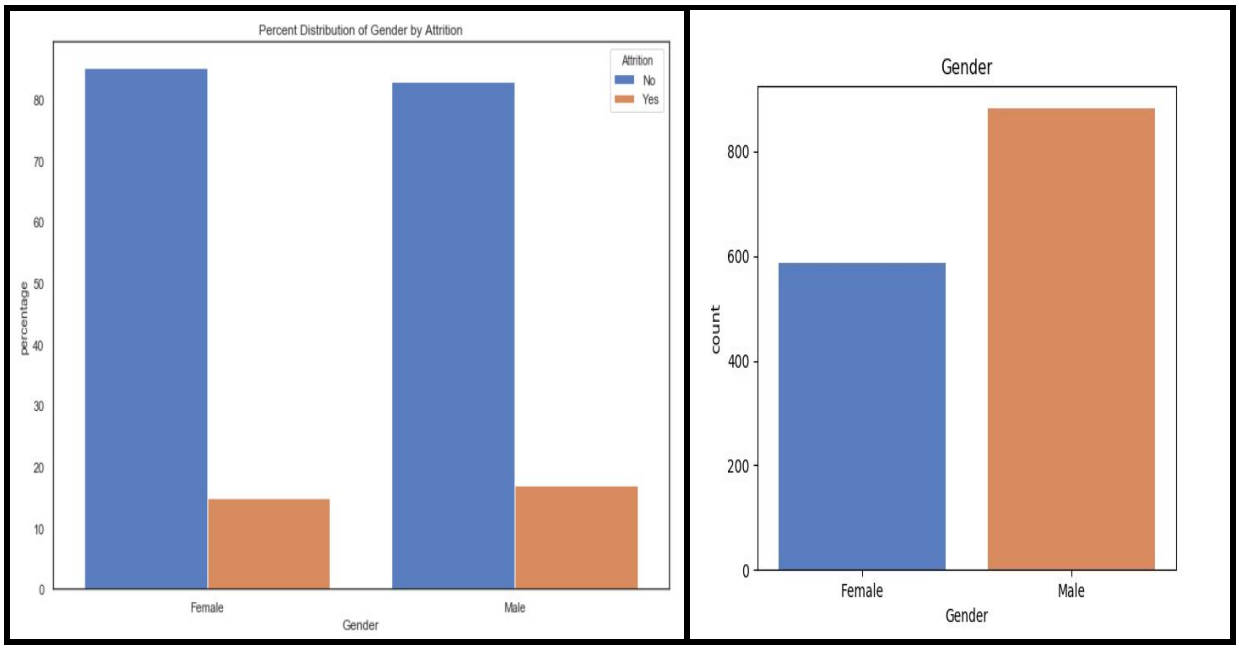


Figure 6: *(a) Gender Count Plot (right), (b) Attrition Rates by Gender (left)*

Finally, to bridge the final gap in our understanding we wanted to know whether attrition was a problem between the two genders. Our analysis shows that the males outnumber the females by 10% (Fig 6a), the attrition rate amongst both is equal (Fig 6b). Through this analysis we were able to proceed forward to create models that will allow us to identify the variables and conditions that encourage employee departures.

Methodology

Through modeling, 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 section is broken down into three subsections that highlight the various collaboration tools, softwares, and algorithms utilized to answer our question.

Collaboration Due to the government guidelines regarding social distancing during the pandemic, the group had to work remotely and rely heavily on technology for collaboration. To ensure progress on the project, a combination of whatsapp groupchat, Github, and a number of google services such as hangouts, docs, and slides, were used. The whatsapp groupchat permitted the group to stay in constant communication and platformed scrum meetings while google hangouts enabled the group to conduct more check-ins such through sharing computer screens. Since the work for each deliverable was equally divided up equally amongst each team member, Github housed the collaboration regarding the coding effort, and Google docs and Google slides housed the collaborated effort for the report and the presentation, respectively. A high-level breakdown of effort is as follows:

- **Code:** Each member of the team was assigned the construction of a model previously agreed upon as a team, and member one was responsible for compiling it
- **Report:** Each member of the team was assigned a section of the report to compose, member two was responsible of compiling it
- **Presentation Slides and Proposal:** Each member of the team was assigned a section of the presentation and proposal and member three is responsible of compiling it

Additional Tools:

To execute this project, we used the **PyCharm IDE** and **Anaconda** libraries to host the series of models and EDA analyses. Upon running preliminary logistic regression, we were able to determine the four variables eliminated only increased noise and did not have an impactful contribution to our analysis. Additionally, to confirm this, we conducted EDA. Once we decided our dataset was noise free, we scripted the models using Conda libraries in PyCharm. To build the skeleton of our GUI, we used a wireframe called **balsamiq**. A website wireframe allows us to layout the blueprint which would be further developed to host a user-friendly interface displaying our models.

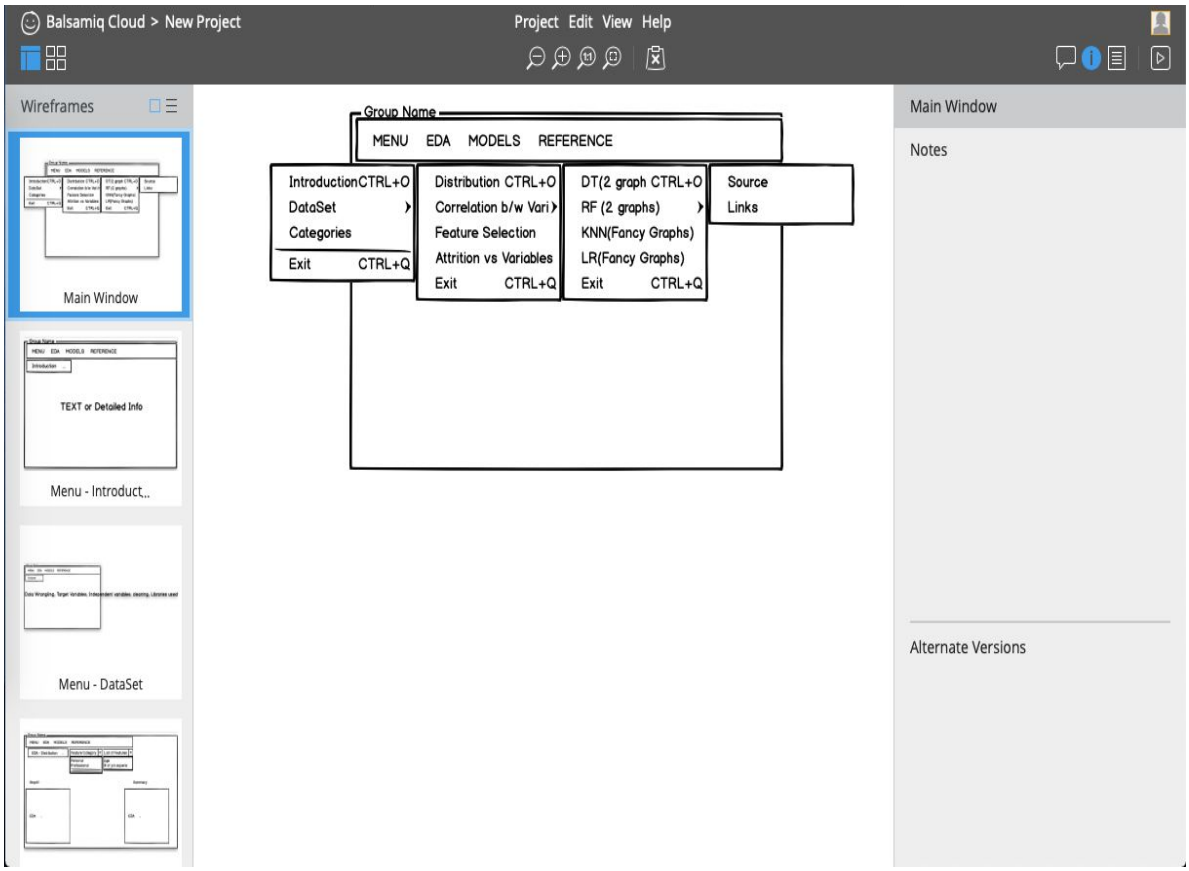


***Figure 7**: balsamiq wireframe*

Results

In the era of big data, machine learning algorithms play a crucial role in that if you train the models, they have the capability to find insights in large dataset, which are often exhausting for humans. For this project, we will use logistic regression, random forest, decision tree and K-nearest neighbor to conduct our analysis. The 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. The following algorithms were used in our analysis:

- **Logistic Regression** is used to predict the probability of binary responses on values of explanatory variables. Given the results of our model, we believe that all variables in this dataset contribute to the employees attrition decision. We incorporated a calibration curve, to assess the accuracy of the model.

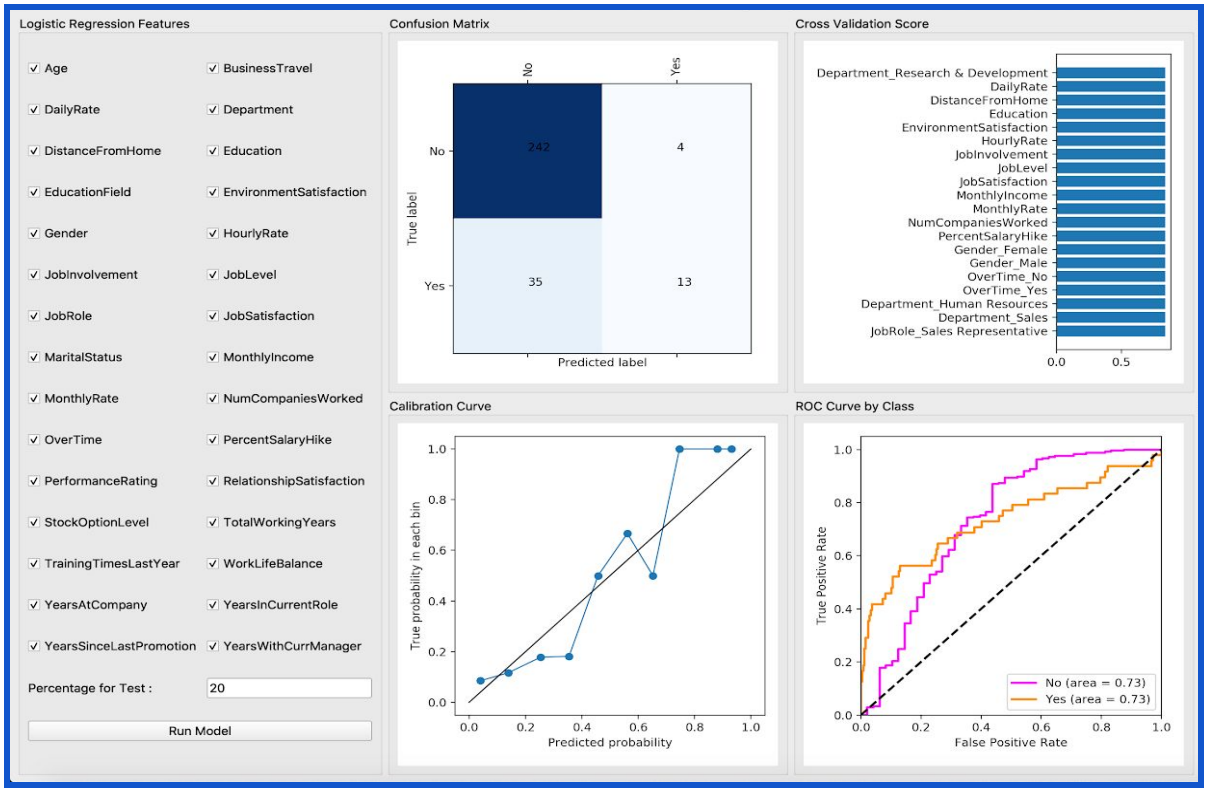


***Figure 8**: Logistic Regression Output*

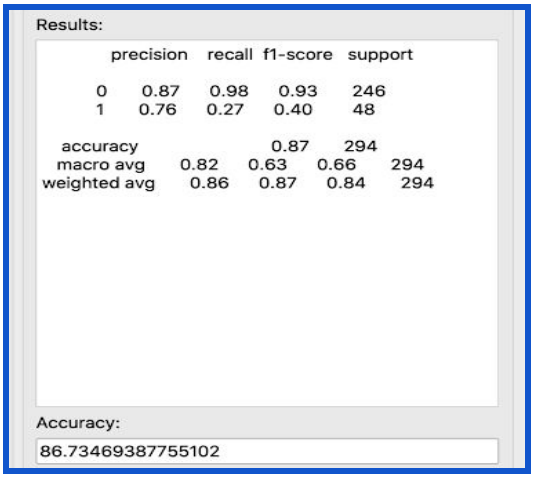
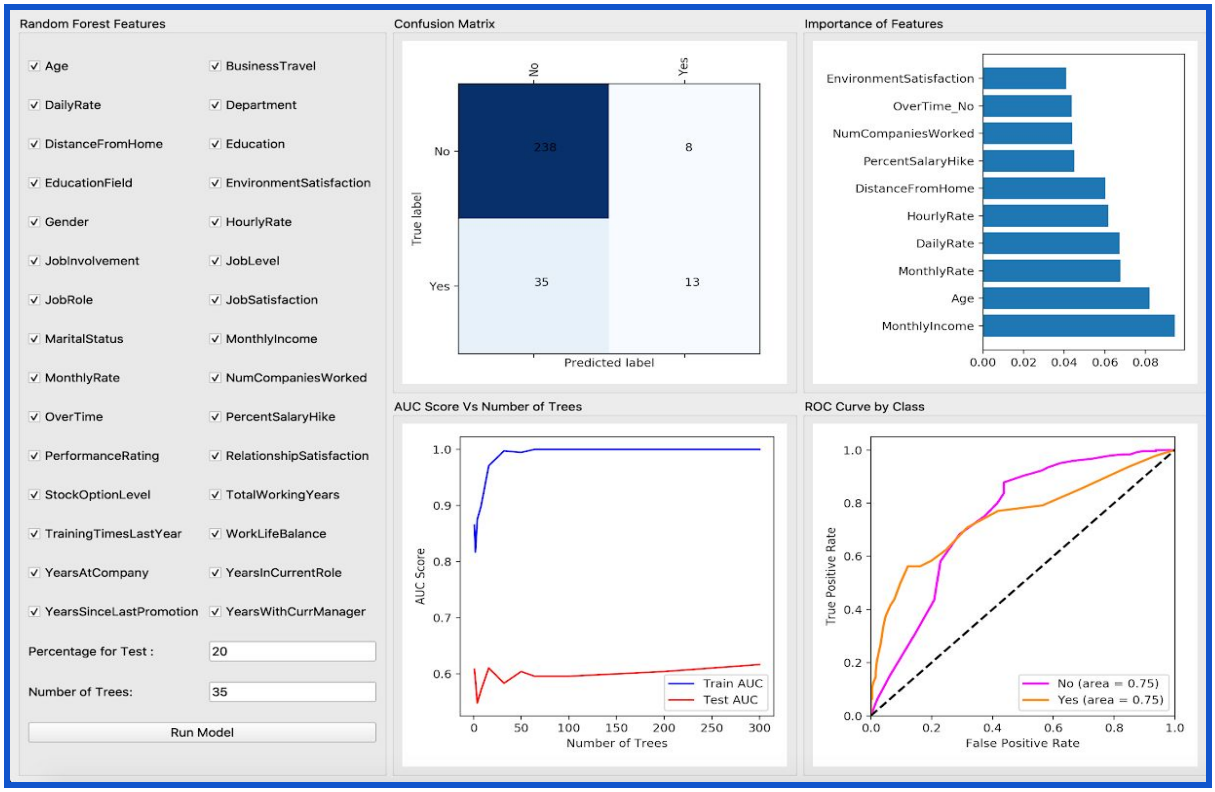
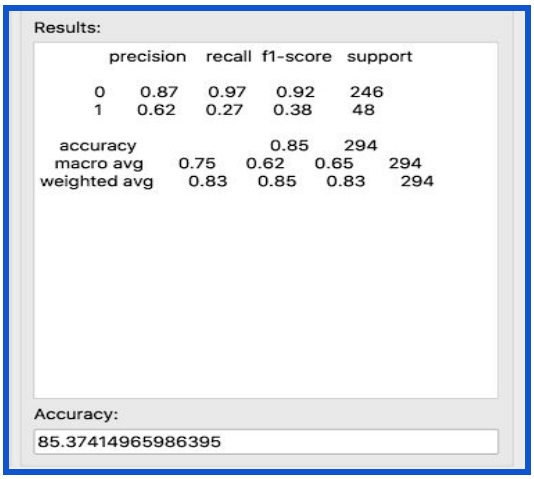


***Figure 9:** Logistic Regression Accuracy Score*

- **<u>Random Forest</u>** algorithms come with a number of advantages. Through our random forest model, we were able to measure how much the accuracy decreases if the variable is excluded. Additionally, we were also able to examine any decreases of Gini impurities when a variable is chosen to split a node.

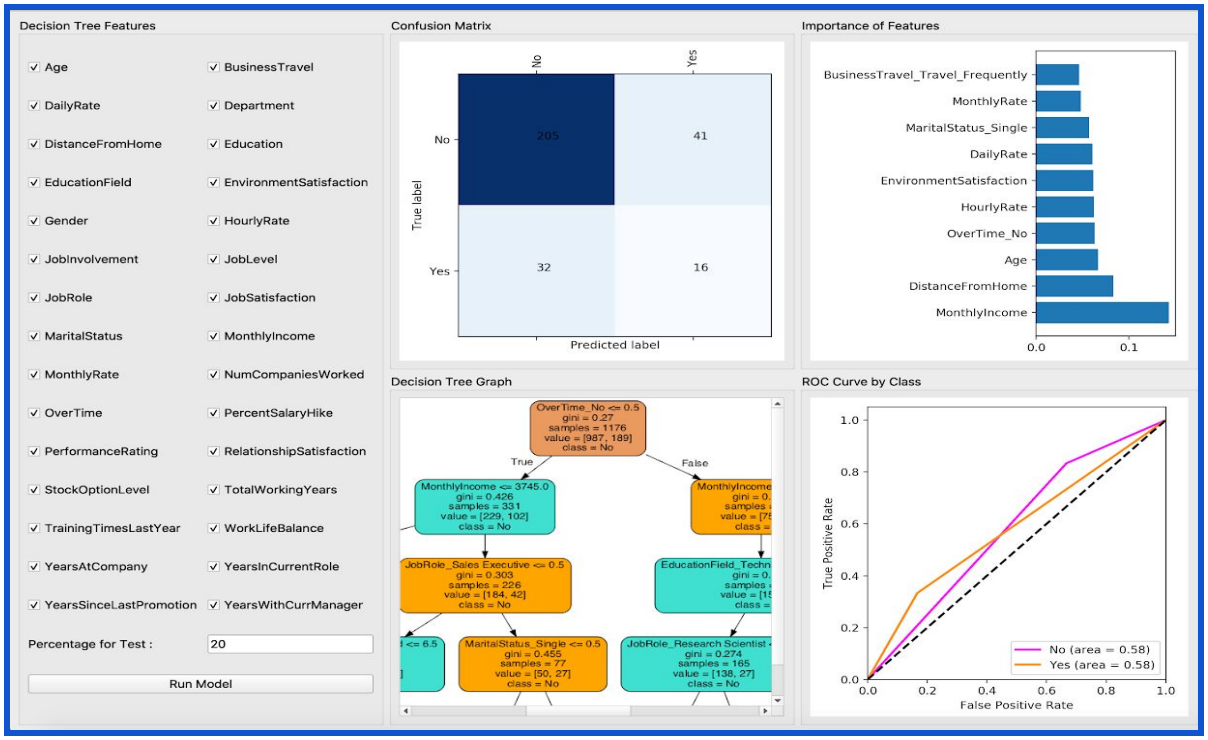


***<u>Figure 10</u>**: Random Forest Output*

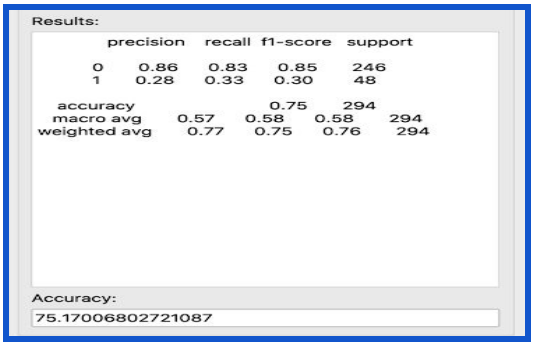


***<u>Figure 11</u>**: Random Forest Accuracy Score*

- **<u>Decision Tree:</u>** enable a quick comprehensive analysis on the outcome of a decision. The advantage of decision trees is that missing values do not make a considerable impact on the outcome. The accuracy score of 75% tells us that the model is fairly reliable.

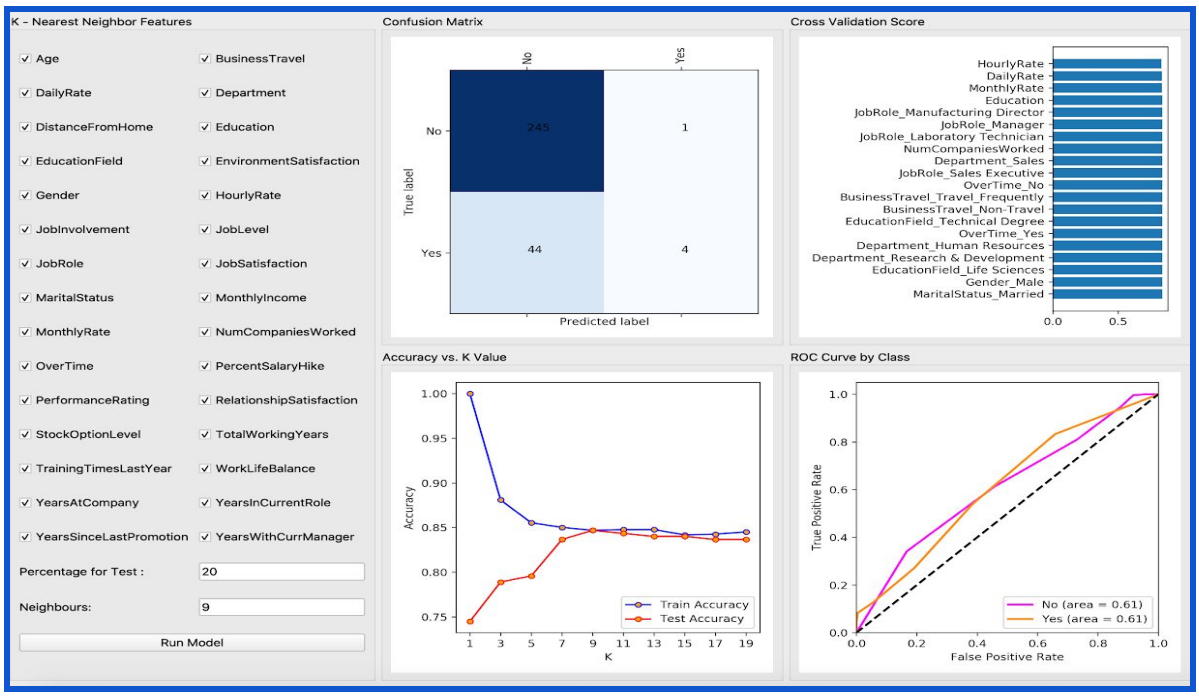


***<u>Figure 12</u>**: Decision Tree Output*

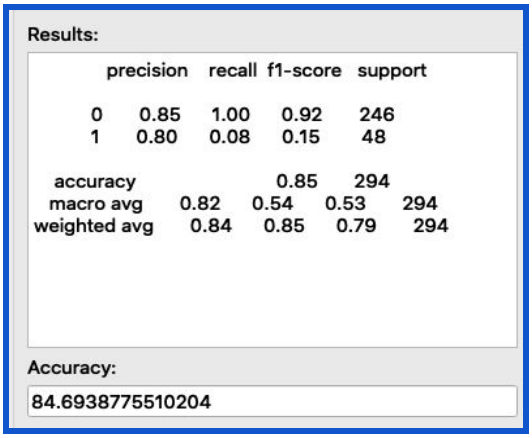


***<u>Figure 13</u>**: Decision Tree Accuracy Score*

- **<u>K-Nearest Neighbors</u>** is often the simplest classification machine learning algorithm to implement. The model essentially will graph the data and uses the k classifier to distinguish the number of data points closest plotted to determine its classification. For example, the most optimal K we found is set to 9, giving us an accuracy score of 85%. This means the model believes, based on 20% of our dataset being used to train the model, that clustering data based on the nine closest points plotted will allow us to cluster our data optimally. As we increase the value of k, the model seems to fit better on the test data whereas if we look at the values from k=1, the model seems to be overfitting.



***<u>Figure 14</u>**: KNN Output*



***<u>Figure 15</u>**: KNN Accuracy Score*

Sr. No	Model Name	Accuracy	Benefits	Trade-offs
1	Decision Tree	76.5	Significant variables are automatically selected.	May give lower accuracy without pruning.
2	Random Forest	86.4	Good Accuracy	Minimal / lower interpretability
3	Logistic Regression	86.7	High interpretability Significant variables can be easily identified	Lower accuracy without pruning.
4	KNN	84.7	Simple Implementation; No Training	Space and time taken is more

Table 1: *Final Accuracy Scores*

When comparing accuracy scores in Table 1, logistic regression quickly rises as the most reliant and accurate model.

Conclusion

Through our analysis we learned that the branch of IBM we studied, has no disparities in demographics. In other words, we sensed no type of discriminaiton in the candidates hiring process. The logistic regression confirmed that the variables identified undoubtedly contribute to the strength of our analysis. We believe the turnover is a result of the level of satisfaction with the employees type of work as we saw higher attrition rates within the sales representative and lab technician positions. As with any data science project, there is more work to be done to mitigate any biases and derive additional analysis. Through this project, we were able to identify the conditions under which employees leave, we would be interested in understanding how different recruitment strategies find long tenured employees.

References

“The Hidden Cost of Employee Turnover.” *Alvernia Online*, Alvernia University, 20 Feb. 2019, online.alvernia.edu/articles/cost-employee-turnover/.