Introduction

Problem

This project sought to better understand employee behavior, specifically what would attribute to an employee's likelihood to leave or stay at a company. Insight into this behavior can support and drive business decisions to attract and retain talent, which is more crucial than ever in such a shifting labor market. Companies hold significant amounts of data about their employees, from demographic details such as their age, marital status, and ethnicity to workforce information such as their time with company, department, and position. Tapping into this data allows companies to create a customized and unique approach that best suits their needs.

To gain access to this sensitive data, stakeholder approval is required. The most logical approach to getting approval for analysis would be to approach this problem from a cost perspective. Labor analysis is an integral part of any company's strategy. It is also one of the most cost-heavy factors of any other potential strategy. By retaining internal talent, companies save money in the long run. Attrition and hiring external talent are costly, not only in a fiscal sense but timewise as well. Onboarding, training, and getting a new employee up to task are all incredibly time-consuming. On the flip side, this project can also be looked at as an employee sentiment analysis in that this approach can better support the employee experience if stakeholders gain a deeper understanding of the external and internal factors that impact an employee's decision to stay or leave the company. By pitching this analysis from both perspectives, the project satisfies multiple requirements that align with business needs and development.

Once approval for the project is obtained, ideally, the data would be obtained internally from employee records. Given the nature of this academic project, the data was obtained from Kaggle, from a fabricated source meant to mimic a real-world company's data. Ethically, this type of data should not be available for public analysis, as employee data is considered sensitive information and contains multitudes of personal identifying information even if unique identifiers like employee ID are withheld from the data. Realistically, if an analysis of this scale were to be performed, approval from HR business partners, legal, privacy, and key stakeholders would be required to best approach sensitive information and avoid any potential pitfalls of bias when evaluating complex employee data like demographics.

Details of Milestones

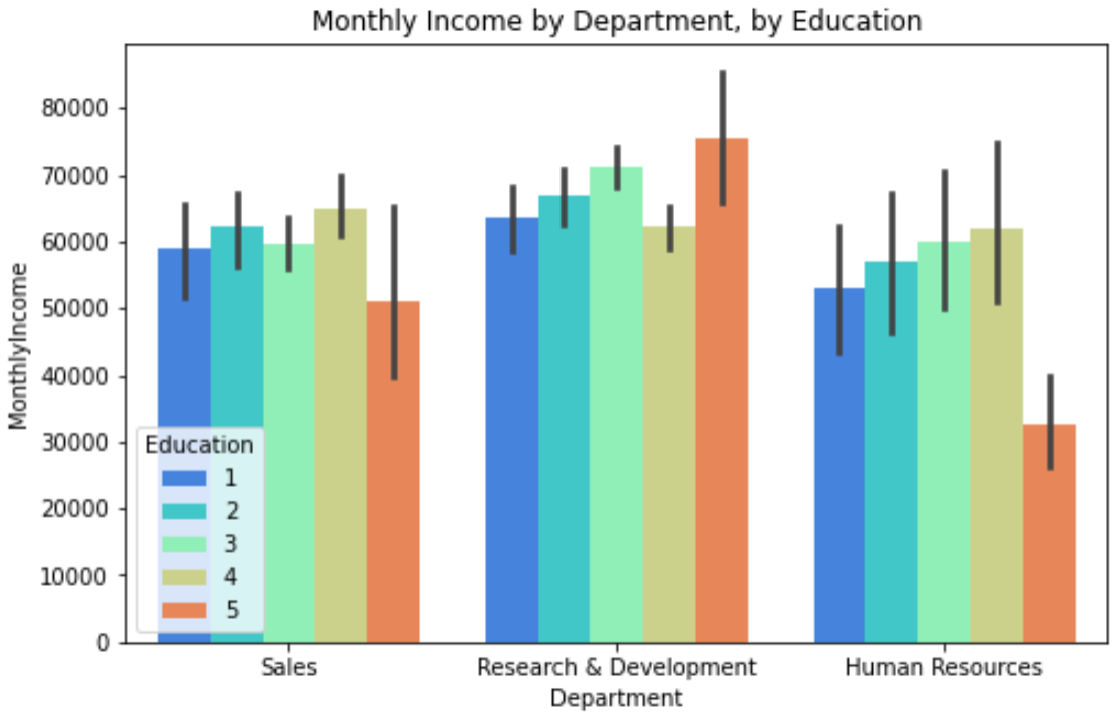
Milestone 1

The first milestone focused on familiarization with the dataset. This dataset was fabricated but had more than 4,000 unique rows of data, each representing an employee. Given the fabricated nature of the data, there were not many actual formatting issues present, such as missing fields, incorrect value formatting, and so forth. This was not surprising given the type of data evaluated as employee records lean towards comprehensive.

Early exploratory visuals did not show any distinct cues to look at further, such as heavy outliers or unexpected trends in the data. The most interesting visual was the employee salary within a department by education level. Education level was represented numerically, as the creator of the dataset provided a dictionary for meaning. Level 1 education represented below college, and level 5 represented PhD. As seen in figure 1, Research and Development, on

average, had the highest monthly income, while Human Resources, regardless of education level, on average, had the lowest monthly income. Overall, there wasn't a huge fluctuation in this dataset. However, given the private nature of salary disclosure, it's difficult to ascertain whether this was due to the data being fabricated, the size or location of the company, or the population sampled.

Figure 1



Milestone 2

Given the low volatility of categorical data found in Milestone 1, Milestone 2 introduced a table join between the original dataset and employee sentiment data. This introduced additional qualitative responses from the employees that included job satisfaction, environment satisfaction, and work-life balance. This can be especially challenging to navigate, as employee sentiment data can be influenced by a multitude of external and internal factors. Implicit understanding

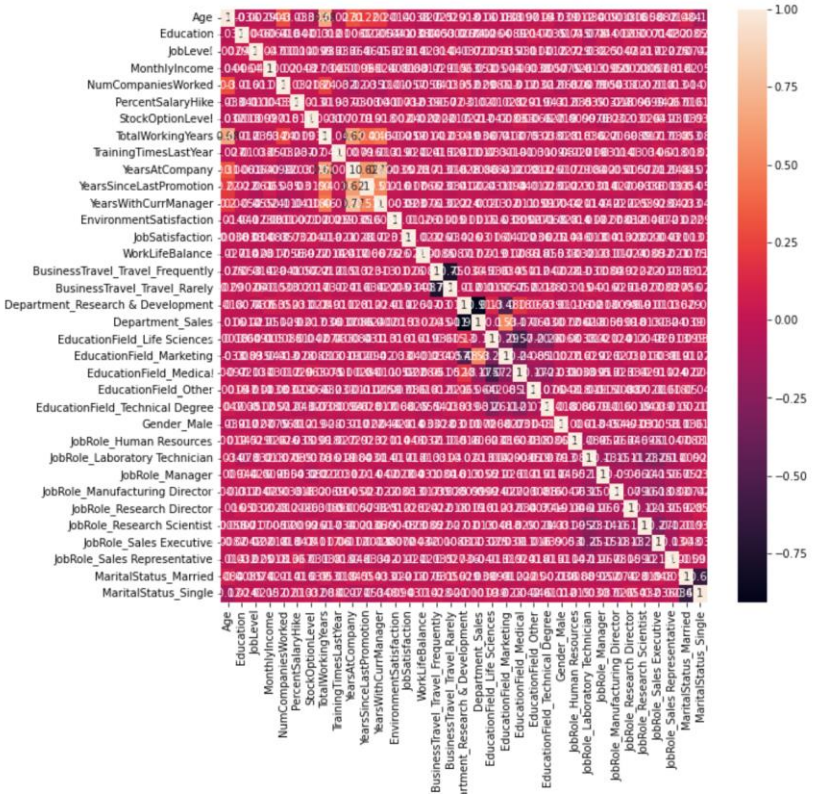
needs to be discussed with the key partners (HR, Legal, etc.), as there is an additional chance for bias to impact any recommendations. This dataset was fortunate enough to one-to-one map employee sentiment data to employee data, but in practice, this would be unlikely as a survey would need to be voluntarily taken with personal identifying information provided by the employee. This could introduce bias, as employees may not freely express their feelings in this scenario. However, since this data was fabricated, the values were taken at face value. Deeper ethical considerations would be taken into account before action in a real-world scenario.

Additional work to clean the dataset was done, such as checking for new null values, removing any potential duplicates, and dropping irrelevant fields.

Milestone 3

In this milestone, work to the actual model began. At the time of this milestone, work was done using a ride regression model, as the data involved so many categorical fields with several values within each field. Without a deep understanding of what fields may incorrectly mislead the analysis or hold dead weight, all fields were taken into account. Results were poor with insignificant improvements made by adjusting the alpha values within the model. As seen in figure 2, a heat map was built to try and decipher which fields held less value, but there was no clear answer available.

Figure 2



Final Milestone

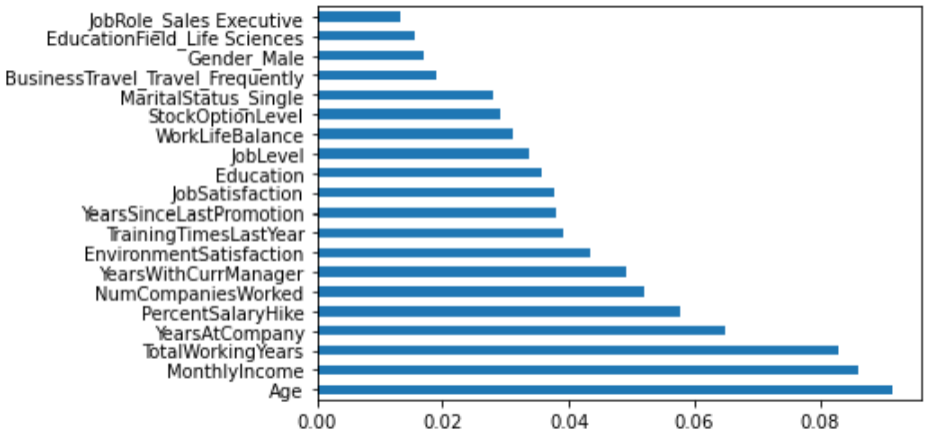
After receiving feedback and guidance from Milestone 3, the Final Milestone sought to build a classification model. While Milestone 3 converted Attrition to binary values, a regression model was still not the correct choice and would be unlikely to return any better results even if additional fields were dropped. Looking at the classification options, a decision tree model seems to be the best fit for the data given its flexibility and low error rate.

The new decision tree classification model returned a 99% accuracy, which was a huge boost in comparison to the ridge regression model used in prior milestones. The predictors also

indicated categorical fields that had the highest impact on attrition, which can further promote future research projects.

Conclusion

Figure 3



The model building in Milestone 3 and the Final Milestone was incredibly rewarding and challenging. Determining the right model for the data was a new obstacle, but hopefully the final model returned enough relevant data to inspire next steps forward. As seen in Figure 3, the fields with the most impact on attrition were age, monthly income, and total working years. Looking further at the data, attrition by age was actually more concentrated in younger employees, despite initial hypotheses that this would be workers closer to retirement age. Next steps for this area of work would involve diving deeper into the categorical fields with the most impact on attrition to develop a better understanding of employee behavior. This could include diving into a marketing-style mindset of creating personas of employees to see what unique groups exist within the company. If younger employees don't feel satisfied with their income or environment, what changes can be made internally? Would the costs be worth the investment in talent?

Overall, the model is ready to deploy. While it can lead to future analyses and projects, this scope of work did not fully answer the broader question of employee attrition and talent retention. However, it did build a strong foundation within the existing employee data and set the understanding that insights can be made from existing data.