Allison Forte

September 11, 2022

DSC 630

Milestone 2: Data selection and project proposal

# Project overview

For this project, I will be working with data relating to employee attrition. I plan to use the data set to analyze factors that may predict employee attrition. Many companies struggle to retain employees and with the current lack of workers looking for jobs, retention is more important than ever.

Several advantages may be possible if a company can identify which employees are likely to stay and which are likely to leave. First, companies can focus efforts on employees likely to leave. Second, companies may have insight into why certain employees stay and use that information to entice others to stay. Finally, companies may be able to alter hiring practices to hire employees who are likely to stay.

In order to complete this project, I will look at factors relating to longevity with the company. This assumes I have data from employees still with the company as well as data from employees who have already left. Knowing the start date for all employees and exit date as applicable is essential. Demographic information about each employee when hired (including age, gender, ethnicity, address (and therefore commute distance and time) and education level as well as hiring salary) will be important when considering which factors influence the likelihood of leaving.

Employment data will also help predict the likelihood of leaving. These factors could include average time in first role, promotion data, raise data, and performance reviews. This type of data would help with predicting tenure based on performance in the role.

Both the demographic data and the performance data could be looked at independently as well as combined. It would be interesting to find predictive factors from the demographic data that could then be coupled with performance in the first few months to get a more granular prediction of an employee’s likelihood of leaving.

I hope to use this project and the methods in it to work with data from my own company to explore ways of improving our attrition.

# Data sources

I will use a data set from Kaggle with information about IBM attrition rates. This data set was created by IBM data scientists. It contains information about more than 2000 fictional employees.

Each employee is identified by an employee number and has 34 additional attributes including age, gender, education level, education field, total working years, and number of companies worked for. Employment data includes job role, department, job level, job involvement, job satisfaction, years since last promotions, years with current manager, years in current role, years at company, training times last year, stock option level, standard hours, performance rating, percent salary hike, overtime, hourly rate, monthly rate, monthly income, daily rate, and amount of business travel. Other factors noted include work life balance, marital status, relationship satisfaction, environmental satisfaction, and distance from home.

Link: <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

# Additional questions

## What types of model or models do you plan to use and why?

I believe I can use a classification model, more specifically, a churn model to predict employee turnover in a similar way to how companies predict customer churn. By adjusting the inputs, I should be able to place employees into two categories based on whether they are likely to leave or not. I hope to generate a probability for each employee to determine how likely it is that they belong in their category.

## How do you plan to evaluate your results?

For this project I believe it will be most impactful to divide the data into a training set and a test set. Once I use a training set to create a model, I will then test the model on the test set of data. Looking at the accuracy of the model will indicate whether it is better than chance. Using several variations of training sets will allow me to look purely at demographic information, purely at employment data, or at a combination of all data available. Comparing the accuracy of these different models will show whether specific factors are a better indicator than others.

## What do you hope to learn?

With this project, I am hoping to learn how to use employee data to predict whether an employee is likely to leave the company. I want to know how to properly prepare the data and how to use predictive models on basic HR data. I do not expect to find any predictions that are useful due to the data being fictional but understanding the methods to create predictions will allow those methods to be used on other, real-world data sets.

The goal of this project is to create a model that can be applied to any company's HR and performance data to create a reliable prediction of an employee’s likelihood to leave the company.

## Assess any risks with your proposal.

Upon first inspection, this data set, while complete in many aspects, may be lacking in performance data. This data set has a performance review attribute, but the only values are 3 and 4. This may not provide the insight into performance’s effect on attrition that I am hoping to explore. Additionally, given this data set has been created, any findings will likely not be valid. The benefit of this analysis will be an understanding of the models that can be used and the methods for preparing data for these models rather than the predictions themselves.

## Identify a contingency plan if your original project plan does not work out.

Kaggle has several data sets that could work for this project. I would like to use the one I specified as it has a lot of detail about each employee and their employment and education history. I suspect that this will lead to a fairly accurate predictive model. However, there is another, seemingly simpler data set that was created for the purpose of predicting attrition. If I struggle with the data I choose, I could switch to the other data set so I can still test my methods. I may not be able to look at as many factors, but I would still be able to create a model.

I might need to look only at demographic data with my chosen data set given the limited amount of performance information.

Back up attrition dataset: <https://www.kaggle.com/datasets/davidepolizzi/hr-data-set-based-on-human-resources-data-set?select=tbl_Employee.csv>

Another alternative back up: <https://www.kaggle.com/datasets/pranavsai98/predicting-employee-status>

# Conclusion

This project is of interest to me because attrition is something my company struggles with in for specific positions. If I can create a reliable model through this project, I can work with our HR department to obtain the necessary data and apply this model to it. If I can uncover predictors of attrition, the company could put effort into the right place to lessen attrition and focus less on hiring and more on retention. I believe this would have an extremely positive impact on my company as well as any other company struggling with attrition.