# Talent Management Attrition Case Study

In this case study, I analyzed a data set with multiple potential features to determine which features were highly telling in attrition rates. Out of the 840 records given, there were 140 that landed in the numerator for attrition rates. In the first two chunks of code, I loaded int the libraries that would be used for this case study and created several plots to determine contributing factors

## Contributing Factors

Some of the most correlated data points in attrition rates were:

Monthly income/Monthly Rate

Job Level

Overtime

Age

## Attrition Predictive Modeling

Once I was able to figure out several data points that were highly correlated with attrition rates, I ran several predictive models (KNN, Naive Bayes, and Random Forest) to predict if an employee would fall in to the denominator for attrition rates.

### Knn Model:

This model performed fairly well in accuracy (84%) and Sensitivity (99.2%), but failed to have high specificity rates. This was concerning as there were a lot of false positives in this modle

### Naive Bayes:

Similarly to the KNN model, the Naive Bayes model performed well in accuracy (85%), but the sensitivity and specificity were opposite of the KNN model. The Naive Bayes model was able to accurately predict the employees who were not in the numerator of the attrition rate, but was very inaccurate in that it predicted a lot of false positives.

### Random Forest:

The random forest model was by far the best model that I ran for predicting attrition rates. It's accuracy (80.8%) was slightly lower than both the KNN and Naive Bayes models, but it's sensitivity and specificity were drastically higher (61.5% and 84.2% respectfully). This meant that this model was able to more accurately predict true positives and true negatives while not sacrificing accuracy for false positives and false negatives

## Salary Predictive Modeling:

I then ran two separate models for predicting salaries for employees. The highest correlating fields were Job Level, Total Working Years and Age. This makes sense as the older you are, the more years you would have worked and your job level is usually higher.

### Random Forest:

When running a random forest model using the fields provided above, I was able to return a RMSE of 1242.69. This was slightly better than the RMSE when using all fields in this model (RMSE of 1278.82)

### Linear Regression:

I then wanted to compare the Random Forest model to a linear regression model, because salary theoretically should be linearly correlated with age, job level and years working. However I found that this models RMSE (1333.81) was slightly higher than the Random Forest model and therefore did not predict salary as accurate.