Human Resources Attrition Rate Project

Boris Tsao

Framing the Problem

As an HR consultant, employers typically come to my firm to ask to do some investigative work into the top reasons why employees are leaving the company.

While it is impossible to have an employee attrition rate of 0% (i.e. nobody leaving the company), employers want to know that they are not:

1. Losing top performers
2. Losing a high proportion of employees who are vital to the business
3. Ensuring that the employer has a stable attrition rate that is not too high

If employers are suffering from any of these 3 (as well as other issues), they may potentially be branded as a "bad" employer and will lose brand reputation from a workforce perspective.

As a data scientist, we will be getting a dataset from an HRIS (human resource information system) datafeed and will be exploring the data there to see what interesting correlations we can find and if we can have a model predict employees that will be leaving and who are staying

Dataset Source

As mentioned in the above section, we will be taking our dataset from an HRIS system. In our case, we got our dataset from a sample company. Typically this is what companies will provide for us for our preliminary analysis. Once we have done exploratory data analysis and able to come up with some insights, we will have more questions and further data points that we may need to explore and merge into our existing file to see if the datapoints are needed.

Exploratory Data Analysis

Here we've done complete some data analysis and found the following generalizations:

1. Younger employees tend to leave the company more – per figure 3 below.

2. Employees in sales tend to leave the company more.

3. Females tend to leave the company more. – per figure 2 below.

4. There does not seem to be a high distinction between high performers leaving the company or staying. – per figure 1 below.

These are all findings that we can present using our business acumen to show that these are relatively normal behaviours in any company. There are no red flags to indicate that this company is performing poorly from an employee retention perspective.

We have also complete a correlation matrix which indicated that we have any highly correlated variables in our data. For our next step, if we had any highly correlated variables, we would have to perform a Principal Component Analysis (PCA) to reduce our variables to ensure that we are not having any highly correlated datasets. Below are the graphs that will show our findings.

Figure 1

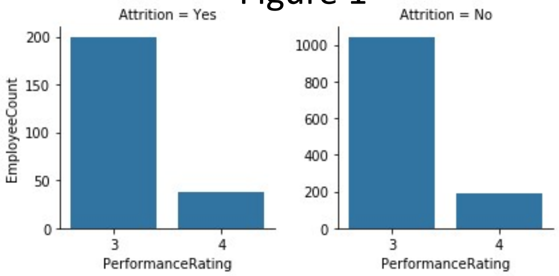


Figure 2

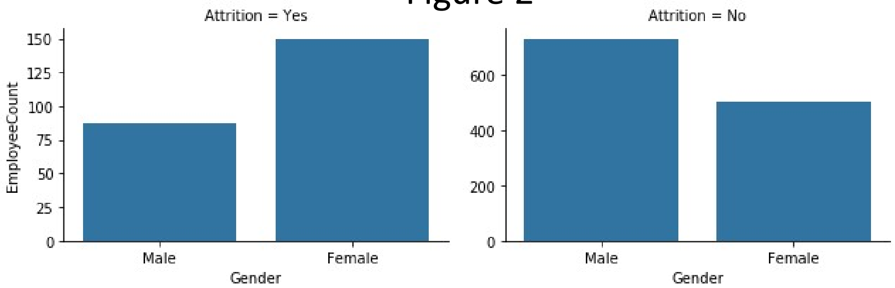
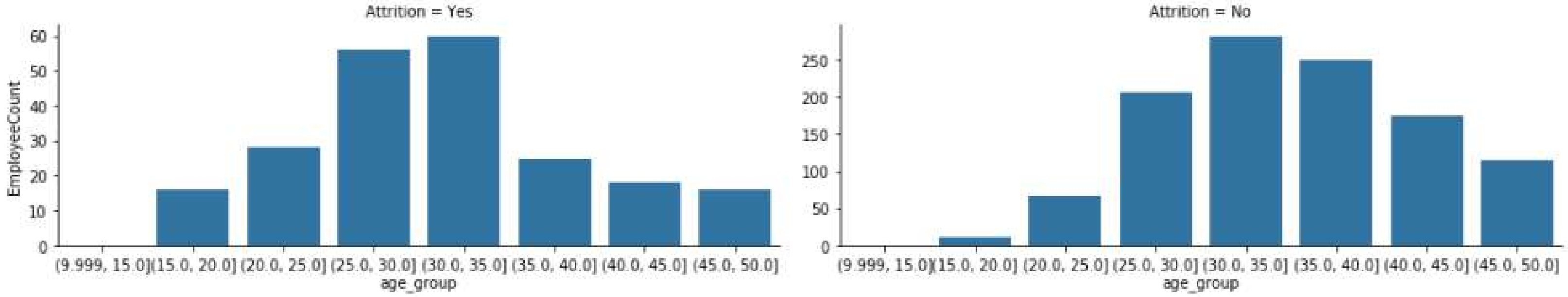


Figure 3



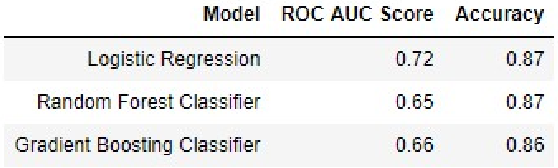
Predictive Model Selection

As we are looking to predict out a discrete variable, we are looking at this from a classification problem perspective. Thus, we will be using the following classification algorithms:

1. Logistic Regression
2. Random Forest Classifier
3. Gradient Boosting Classifier

Note that before we do this, we have to complete an oversampling method as our dataset is heavily imbalanced towards employees staying rather than employees leaving. In order to do this, we will be completing the SMOTE technique (synthetic minority over-sampling technique). The summary is in our "Capstone - Machine Learning" page.

For the performance of our algorithm, we will be focusing on the ROC AUC score and the accuracy. We want a model to be able to predict accurately who is staying and leaving (the accuracy score), but are more focused on the true positive score (ROC AUC score). The full detail of our implementation is in our "Capstone - Machine Learning" section. For now, we will present out the highlights



Model	ROC AUC Score	Accuracy
Logistic Regression	0.72	0.87
Random Forest Classifier	0.65	0.87
Gradient Boosting Classifier	0.66	0.86

We will pick the random forest classifier to hyperparameter tune (despite our logistic regression model having a better AUC curve) as our model was able to predict out a higher number of true positives than every other model.

There are several ways to tune our hyperparameter. For us we used the randomsearchcv method and was able to have a higher AUC score.

Conclusion

We have conducted several steps from EDA to hyperparameter tuning our model. We were able to train our model and returned 87% accuracy. We can use our model now on a monthly or bi-annual basis to see which employees are at risk of leaving and use our model to dig into whether they are overtime eligible and who are top performers to try and persuade them to stay within the company.

Although we were able to predict this out, we can perform more feature engineering to get more out of our data. We can also talk more with stakeholders to interpret our results and gather more data and more intuition on how we can use findings.

Our next steps would be to use our trained model to run through the existing employee dataset, see who are model says are most likely to leave and for our stakeholders to conduct interviews who employees who are indeed at high risk of leaving to see if there is any further insights we can gather from this.

Another step we could do is see if we can get more data on other interesting variables (e.g. # of dependents, employees going through different stages of their life, how much they contribute in their 401(k)). This would give us a better understanding on what other variables go into an employee leaving the company.