HR ANALYTICS: CAN WE PREDICT ATTRITION?

Chris Henson | LaShay Fontenot | Rawini Dias | Chris Fitzgerald | Yikang Wang

Introduction

The HR Analytics dataset consists of employee data for a company. The dataset has 4410 rows and 28 predictor variables. The variable of interest is employee attrition. The goals of this project were to determine the significant predictors of attrition at this company and to build a model capable of predicting employee attrition in the future.

Data Pre-Processing

The dataset was initially cleaned by removing redundant and unusable variables. We identified missing data by using a missingness map and made an informed decision to omit the rows of missing data. Feature engineering was performed on some variables to create more effective variables.

Exploratory Data Analysis

We utilized the ggplot library to perform exploratory data analysis on the dataset. The exploratory analysis did not uncover any striking trends in the data. However, it did yield some useful insights into the data in the way of employee distribution by monthly income, job level, work life balance, etc.

Logistic Regression

Since the variable of interest, attrition, is categorical, we first ran a logistic regression model using all the predictors in the pre-processed dataset. Probabilities greater than 0.5 were classified as ‘Yes’ for attrition. This model yielded an out of sample misclassification error of 0.15. Based on this model, only 61% of the employees predicted to leave the company actually left. We concluded that we needed a more effective model.

LASSO

After running our initial logistic regression for employee attrition with the inclusion of all available predictors, the need to cull our included variables became readily apparent.

After running regressions several times with different training/test datasets, it appeared that the full logistic regression only had on average an accuracy of 1-2% higher than the individual logistic regressions of one predictor. Along with the fact that only a handful of predictors had statistically significant coefficients in the full model, this suggested that several of the included variables were not contributing much overall.

As such, this presented a perfect opportunity to utilize the LASSO method, which presented us with approximately a dozen variables which were not pushed to zero after cross validation through lambda values. While the misclassification error remained approximately the same, the proportion of error found for employees predicted to leave the company decreased significantly, while the proportion of error for employees predicted to remain with the company remained stable.

While this model served as a significant improvement, we still believed that we could achieve greater accuracy (though perhaps at the expense of ease of interpretability).

Random Forest

At this point we had a LASSO model which was more simple than our logistic regression (by dropping 14 variables) but still had an error rate greater than 10% on out of sample data. We wanted the error to decrease so we tried a random forest with 500 trees and the default $\sqrt{p}$ (p= 28). This resulted in an accuracy of 97.6% overall. However, using the output of our LASSO model we were able to drop 14 variables from our model. This created a simpler random forest model using 500 trees and $\sqrt{14}$ variables considered at each split. The accuracy increased to 97.8% overall and of the people who were predicted to leave the company, 96.9% actually left, and of the people who were predicted to stay with the company, 98% didn't leave.

Model Performance Evaluation

Comparing the three different models that were developed, the Random Forest model built off of the LASSO variables comes out on top. This model produced the highest prediction accuracy (.978), while minimizing the occurrence of obtaining a false positive or negative. In fact, the Random Forest model reduced the occurrence of a false negative by a factor of 10 from the original Logistic (.39 vs .031). Assessing out of sample misclassification error, we see that LASSO and Logistic produce the same value of .15. However, if prompted to select the better model out of the two, we would suggest using LASSO. This model reduces complexity by zeroing out many of the variables in addition to cutting the occurrence of a false negative by half (.39 to .18).

Conclusion

Overall, all models attempted performed decently well in predicting employee attrition, with each one performing better than the last. Considering our second objective of the analysis, both the LASSO and Random Forest helped us understand what factors / variables were most important to our prediction accuracy allowing us to reduce the complexity of our models. In the end, we conclude that we can predict attrition most accurately with Random Forest where average daily hours worked, age of the employee, total working years, and job role play the largest role in the prediction.