CANTERRA LOGISTIC REGRESSION

Abigail Hoffman

Georgetown University
Machine Learning I

I. Background

Canterra is a large company that currently employs ~4000 employees. Like many other large companies, given the current job market, Canterra faces a yearly attrition rate of 15 percent. The company believes the turnover is negatively impacting the company's reputation as recruitment of new talent requires additional training and time to acclimate to the company's work environment resulting in delayed projects that do not meet deadlines. Senior management at Canterra has hypothesized that job satisfaction, tenure at the company and longer tenure of total working years reduce employee turnover rates. Additionally, the marketing team is interested in understanding demographic variables that could impact recruiting efforts and effectiveness. Canterra has provided a recent employee data set that explores whether an employee left the company.

II. Executive Summary

The main goal of this report is to understand the driving factors behind attrition at Canterra and model the probability of attrition based on those factors. Given that attrition is understood as a binary variable (yes, the employee left or no, the employee did not leave), logistic regression is determined to be the best model. Senior management's hypothesis held as job satisfaction and total tenure did return as significant factors for attrition; however, years at the company did not factor into attrition. For marketing, the only factor that returned significant was age. Gender and education did not play pivotal roles in turnover for the company. Two additional variables identified as significant were environmental satisfaction and number of training times within the past year. For all variables, a one-point increase resulted in a negative log-odds of attrition at the company. Given the significant predictors in the model, Canterra has an opportunity to offer surveys for an employee listening strategy, a more flexible work-life balance for the office environment, and more training opportunities for cross-departmental interaction.

III. Data Exploration

The data set provided by Canterra offers data for 18 variables (columns) and 4410 employees to explore attrition. The data was missing 73 observations across four variables: TotalWorkingYears, NumCompaniesWorked, JobSatisfaction and Environment Satisfaction. Rather than remove the missing observations, I opted to impute the median value for the variables' total working years and number of companies worked. As job satisfaction and environment satisfaction are categorical in nature (1-4 scale) yet are used numerically for the model, I chose to impute the mean of each category to lessen the bias that using the mode could have attributed.

IV. Methods & Models

All regressions, visualizations and calculations were crafted using packages in R Studio. R studio is an open-source programming language that enables statistical analysis and data visualization.

a. Balancing Data Set for Model

To aid in classification accuracy as a performance measure, I utilized the Rose package in R to balance the classes of 0 and 1 in attrition, where 1 is the employee leaving Canterra. For machine learning exploration, I tested four possibly balanced samples: random over-sampling minority examples, under-sampling majority examples, a combination of over and under-sampling and a sample of synthetic data by enlarging the features space of minority and majority class examples.[1] To select the best data set, I compared models using all variables against attrition with each type of data classification balance and compared the AIC, Akaike Information Criterion. The best data balance was to use the under-sampling method as it has the lowest AIC.

	Under	Over	Over-Under	ROSE(Synthetic)	Original Data Set
AIC	1183.3	5902.2	2575.3	3728.3	2333.3

b. Modeling

Multiple logistic regressions were performed to address the questions posed by Canterra along with my hypothesis of the cause of attrition. For the purposes of this report, the focus is on the final logistic model used in the training and test data sets.

Logistic Regression (Log-Odds)

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + B_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

The final model's significant variables were selected by the stepwise method.[2] The five variables in the regression model below were determined to be a predictor of attrition.

$$\begin{aligned}\mathrm{logit}(Attrition) = \beta_0 + \beta_1 JobSatisfaction + \beta_2 TotalWorkingYears + B_3 Age \\ + \beta_4 EnvironmentSatisfaction + \beta_5 TrainingTimesLastYear + u\end{aligned}$$

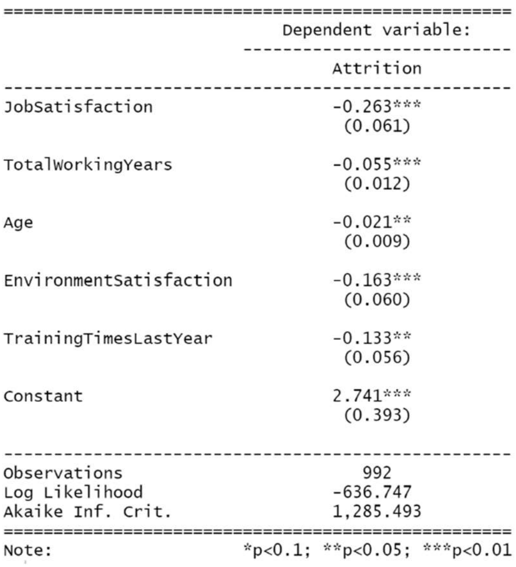
[1] R Documentation

[2] Stepwise regression (bidirectional) is a method that fits the best predictive variables to the regression and is performed by an automatic procedure in R studio.

See appendix part B for all models explored and their results.

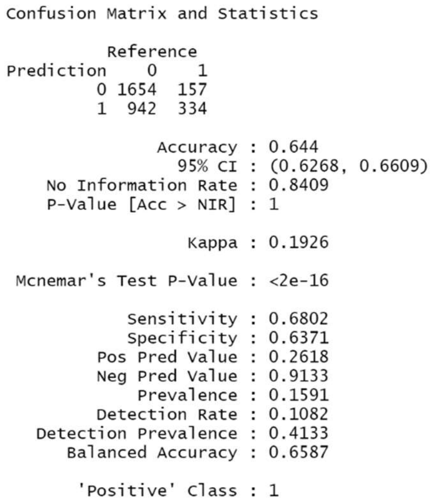
V. Results

Based on the final model, management’s hypothesis was correct that a higher survey result of job satisfaction reduces attrition overall for the company as job satisfaction increases by one-point, the odds of attrition decrease by 26.3 percent. Environment satisfaction was an additional predictor I chose to include as training and hiring new employees to assimilate into a new company takes time and can influence the continued pattern of turnover. This was the second most influential predictor as a one-point increase in satisfaction with the work environment resulted in a 16.3 percent decrease in attrition. In addition to management’s hypothesis, the total training times last year predictor was added to explore a potential opportunity if engaging with work colleagues on new material and topics would decrease turnover. This did result in a 13.3 percent decrease in attrition with just one additional training per year. Age and total working variables were significant for the model; however, the beta percentage is marginal.



```
=================================================
                        Dependent variable:
                    ---------------------------
                             Attrition
-------------------------------------------------
JobSatisfaction              -0.263***
                              (0.061)

TotalWorkingYears            -0.055***
                              (0.012)

Age                          -0.021**
                              (0.009)

EnvironmentSatisfaction      -0.163***
                              (0.060)

TrainingTimesLastYear        -0.133**
                              (0.056)

Constant                      2.741***
                              (0.393)

-------------------------------------------------
Observations                    992
Log Likelihood               -636.747
Akaike Inf. Crit.            1,285.493
=================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

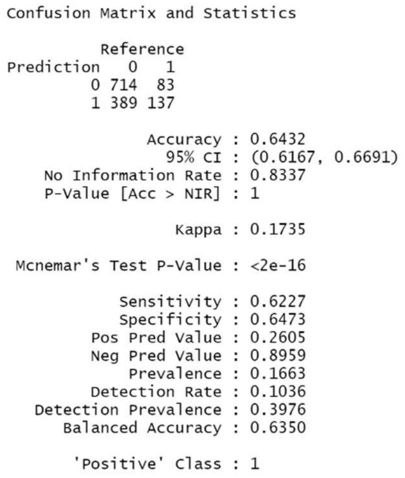
a. Model Performance



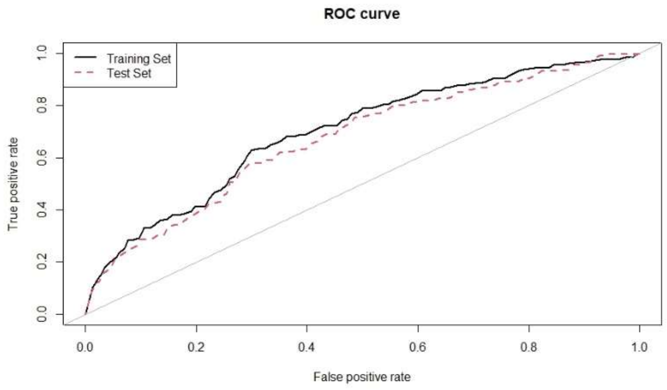
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1654  157
         1  942  334

               Accuracy : 0.644
                 95% CI : (0.6268, 0.6609)
    No Information Rate : 0.8409
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1926

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6802
            Specificity : 0.6371
         Pos Pred Value : 0.2618
         Neg Pred Value : 0.9133
             Prevalence : 0.1591
         Detection Rate : 0.1082
   Detection Prevalence : 0.4133
      Balanced Accuracy : 0.6587

       'Positive' Class : 1
```

The main goal of this project is to model the probability of attrition for Canterra and to be able to apply these predictions to future data sets for the company. The following demonstrates the final model’s ability to predict attrition given the training data. The model resulted in a 64.4 percent accuracy rate and is significant at the 95 percent confidence level. The model’s sensitivity, power, to detect attrition cases is 68 percent when the threshold is 0.5, or 50 percent. Other notable factors include the prevalence rate of attrition occurring in the training set at 15.91 percent and the model’s detection rate of 10.82 percent. The model’s AUC, area under the curve, is 0.698.

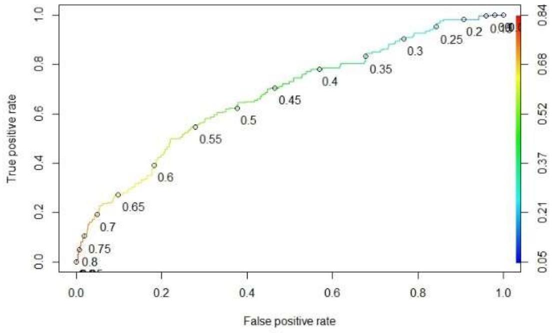
Although the model's performance declined slightly when applied to the test data set, the quality of the model was upheld well. The model resulted in a 64.32 percent accuracy rate and is significant at the 95 percent confidence level. The sensitivity or the power of the model to detect attrition cases is 62 percent when the threshold is 0.5. The amount of attrition occurring in the test set is 16.63 percent and the model's detection rate is 10.32 percent. The test set model's AUC is 0.672. There is a 67.2 percent chance that the model can differentiate between a positive class and a negative class. The graph below compares the training set's AUC to the test set's AUC. Both are relatively close in nature and retain similar shapes.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 714  83
         1 389 137

               Accuracy : 0.6432
                 95% CI : (0.6167, 0.6691)
    No Information Rate : 0.8337
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1735

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6227
            Specificity : 0.6473
         Pos Pred Value : 0.2605
         Neg Pred Value : 0.8959
             Prevalence : 0.1663
         Detection Rate : 0.1036
   Detection Prevalence : 0.3976
      Balanced Accuracy : 0.6350

       'Positive' Class : 1
```



The model's performance slightly decreased in efficiency; however, the test set had a larger value of observations, 1,323.[3] Overall, the model performed well. To better visualize the model's performance in the training and test set, the two graphs below show the models fit in true positive rate against false positive rates. Both models are curved towards true positive rate detection. Another visual below shows the area under the curve with the attrition y values of varying levels.



[3] As the under-sampling technique produced the lowest AIC, this did reduce overall total of training set observations to 992.

VI. Recommendations

Canterra has a multitude of opportunities for improvement to reduce attrition and increase employee satisfaction. After statistical analysis, job satisfaction was the most significant in magnitude as a predictor to reduce attrition. The happier the employee is with the job, the less likely they are to leave. One possibility is to use a network like Pulse Surveys on a quarterly basis to understand employee well-being, departmental pain points and the overall atmosphere of the company. For management, this serves as an effective employee listening strategy.

Another area that arose significant was satisfaction with work environment. Depending on what the company can offer, allowing hybrid workdays with flexible schedules that empower workers to set their days and schedules could positively influence retention. People tend to enjoy a less formal work environment. In general, the company could employ a *how we work* framework that allows varying levels of business wear for what the individual's day looks like.

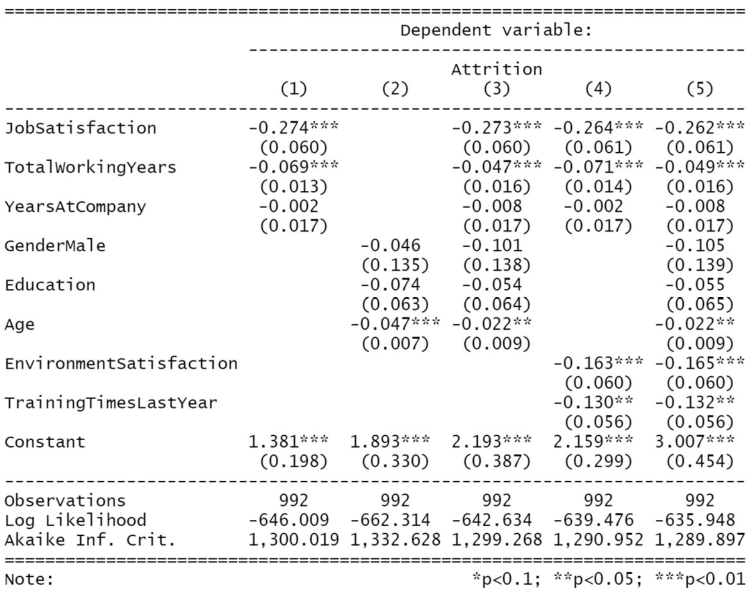
Finally, trainings proved to be largest margin of opportunity for Canterra to reduce turnover rates. Often, trainings allow cross-departmental development. People may be able to meet other employees who can collaborate or support them when a difficult task arises. Another idea is offering department shadowing, where someone may be interested in another department's work and can spend a week with that department to see if it is a good fit. This could reduce turnover as employees may stay within the company rather than seeking external opportunities.

Appendix[4]

A. Description of Data Variables[5]

1. Age (in years)
2. Attrition: Whether the employee left the previous year (Yes, No)
3. BusinessTravel: How frequently the employees travelled for business purposes last year
4. DistanceFromHome: Distance from home to location of work (in miles)
5. Education: Education (1 ='Below College', 2= 'College', 3= 'Bachelor', 4= 'Master', 5= 'Doctor')
6. EmployeeID
7. Gender (Female, Male)
8. JobLevel: Job level at company (scale of 1 to 5, level 1 is lowest and 5 is highest)
9. MaritalStatus: Marital status of the employee (Single, Married, Divorced)
10. Income: Annual Income (in $)
11. NumCompaniesWorked: Number of companies they worked at previously
12. StandardHours: Standard hours of work for the employee
13. TotalWorkingYears: Total number of years the employee has worked so far
14. TrainingTimesLastYear: Number of times training was conducted for this employee last year
15. YearsAtCompany: Total number of years spent at the company by the employee
16. YearsWithCurrManager: Number of years under current manager
17. EnvironmentSatisfaction: Satisfaction with Work Environment (1= 'Low', 2= 'Medium', 3= 'High', 4= 'Very High')
18. JobSatisfaction: Job Satisfaction (1= 'Low', 2= 'Medium', 3= 'High', 4= 'Very High')

B. Model Summary of All Models Before Stepwise Selection



```
=========================================================================
                                       Dependent variable:
                        -------------------------------------------------
                                            Attrition
                           (1)       (2)       (3)       (4)       (5)
-------------------------------------------------------------------------
JobSatisfaction         -0.274***           -0.273*** -0.264*** -0.262***
                         (0.060)             (0.060)   (0.061)   (0.061)
TotalWorkingYears       -0.069***           -0.047*** -0.071*** -0.049***
                         (0.013)             (0.016)   (0.014)   (0.016)
YearsAtCompany           -0.002              -0.008    -0.002    -0.008
                         (0.017)             (0.017)   (0.017)   (0.017)
GenderMale                         -0.046    -0.101              -0.105
                                   (0.135)   (0.138)             (0.139)
Education                          -0.074    -0.054              -0.055
                                   (0.063)   (0.064)             (0.065)
Age                               -0.047*** -0.022**            -0.022**
                                   (0.007)   (0.009)             (0.009)
EnvironmentSatisfaction                               -0.163*** -0.165***
                                                       (0.060)   (0.060)
TrainingTimesLastYear                                 -0.130**  -0.132**
                                                       (0.056)   (0.056)
Constant                1.381***  1.893***  2.193***  2.159***  3.007***
                         (0.198)   (0.330)   (0.387)   (0.299)   (0.454)
-------------------------------------------------------------------------
Observations               992       992       992       992       992
Log Likelihood          -646.009  -662.314  -642.634  -639.476  -635.948
Akaike Inf. Crit.       1,300.019 1,332.628 1,299.268 1,290.952 1,289.897
=========================================================================
Note:                                         *p<0.1; **p<0.05; ***p<0.01
```

[4] Hoffman_GithubP2

[5] Canterra Employee Data