Project 2: Canterra Logistic Regression

OPIM 602 – Machine Learning I

Dr. Sudipta Dasmohapatra

22 November 2021

Jed Raynes

Introduction

Canterra is a large organization with around 4,000 to 5,000 employees and experiences attrition of approximately 15% per annum. High level of employee turnover leads to delayed projects which have impacts on customer deliverables. Additionally, to replace former employees, the company must invest time and capital to train newer employees to cover the responsibilities of former employees. The company has engaged me to analyze an employee dataset which includes various job characteristics, employee demographics, and whether the employee left the organization within the past year. My findings indicate that job satisfaction and the employee's total years in the workforce are the leading indicators of attrition or lack thereof.

Scope, Approach, and Assumptions

The dataset provided consists of 4,410 employees, 17 predictor variables, and one dependent variable. I was engaged to analyze the employee data and provide insights based on the analysis. Given the binomial outcome variable, I used a logistic regression model to predict attrition. A threshold of 0.05, or 5%, was used to determine statistical significance. Management hypothesized that job satisfaction, years at Canterra, and total working years influenced attrition. It was also asked that I consider employee demographics such as gender, the highest level of education achieved, and age. I focused my analysis on these six predictors given they were the focus of management. When predicting attrition, employees with predicted probabilities of 50% or greater are categorized as attrition.

Data Splitting

I split the dataset into a training and testing set with a 70% / 30% split. Given the data provided is for only the past year and to maximize the data available for model training, I did not use a validation set in my analysis. Across the entire dataset, and in line with management's estimate, 16.12% of employees left the company in the past year. Given the imbalance skewed towards employees that stay with the company, a stratified random sampling method was used to maintain a similar weighting of attrition in the testing and training sets.

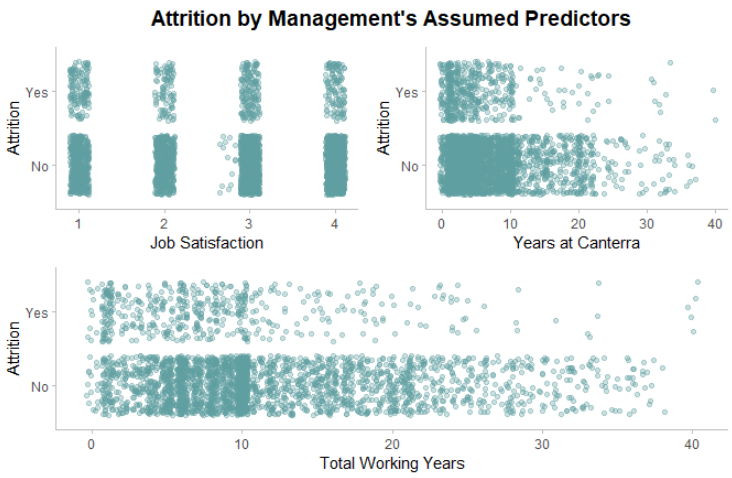
Missing Data

Through the exploration of the training set, I noticed that the following predictors were missing data: number of companies worked (12 missing entries), total working years (6), environment satisfaction (18), and job satisfaction (11). The image below plots the distribution of the four predictors with missing data points.

The total working years and number of companies worked variables appear to be right-skewed and represent numerical data types. Given the skewness, I opted to impute the median value of each predictor for missing data points. The environment and job satisfaction are technically categorical as it is a score between one and four. For the purposes of the analysis, these variables were treated as numerical. For the latter two predictors, I used the mean of each predictor as to not skew the satisfactory ratings of the environment / job given there is no true midpoint of a four-level factor.

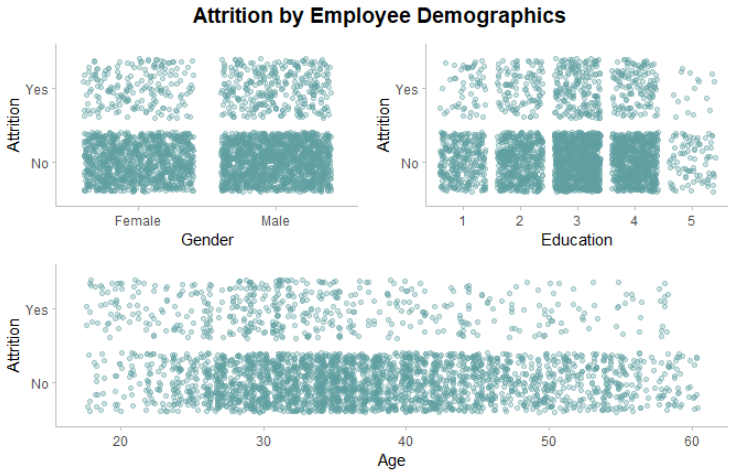
Source: Canterra

Exploratory Data Analysis



Source: Canterra

After the imputation of missing data, a robust exploration of data was performed to assess relationships between predictors and employee attrition. First, I explored management's assumed predictors: job satisfaction, years at the company, and total working years. Job satisfaction responses were primarily three or four, however, while attrition occurred across all satisfaction levels, there tends to be more attrition at the one and three level. Most of the employees have been with the company between one and 10 years and attrition tends to occur with employees with less than five years of tenure. Employees tend to have between five and 10 years of total years of work experience with attrition typically occurring with those between one to two years, five to six years, and ten years of experience. The image to the above displays the density of each predictor separated by attrition and no attrition.



Source: Canterra

In addition to management's hypothesized predictors, I explored the employee demographics of gender, highest level of education achieved, and

age. The company's employee base tends to be slightly weighted towards more males, employees primarily have a bachelor's and master's degrees with few having doctorates, and employees tend to be approximately 25 to 45 years old. The image above displays the employee demographics by attrition.

Data Sampling

As mentioned previously, I used stratified random sampling when splitting my data given the imbalance in the dependent variable. To better predict the probability of attrition, I explored sampling methods to offset the imbalanced classes. Employee attrition, which is the minority class, consisted of 16.2% of the training data before sampling. *Refer to Appendix #1 for each sampling methodology and related impacts on the training set.*

To compare the sampling methods, the Akaike information criterion ("AIC") scores, which estimates prediction error, of a full logistic model (i.e., inclusive of all 17 predictors) was compared. The sampling method with the lowest AIC, the under-sampling method, is used as the basis for training the logistic model.

	No Sampling	Over and Under	Over	Under	ROSE (Synthetic Data)
AIC	2,374.8	3,680.6	6,072.5	1,180.6	3,774.5

Modeling

Four logistic regression models were fit to the under-sampled training data to predict attrition.

Model 1: Management's Hypothesis

$$logit(p) = 1.31 - 0.22(JobSatisfaction) - 0.06(TotalWorkingYears) - 0.01(YearsAtCompany)$$

In the first model, job satisfaction and total working years are statistically significant predictors which indicates that they impact likelihood of attrition. Given the negative coefficients for both, as job satisfaction increases by one or total years of working experience increases by one, holding all-else equal, the log-odds of attrition decreases by 0.22 and 0.06, respectively.

Model 2: Employee Demographics

$$logit(p) = 1.59 + 0.12(GenderMale) - 0.06(Education) - 0.04(Age)$$

In the second model, age is the only statistically significant predictor with a negative coefficient. This indicates that as an employee's age increases by one year, holding all-else equal, the log-odds of attrition decreases by 0.04.

Model 3: Management's Hypothesis and Employee Demographics

$$logit(p) = 1.70 - 0.22(JobSatisfaction) - 0.05(TotalWorkingYears) - 0.01(YearsAtCompany) + 0.09(GenderMale) - 0.06(Education) - 0.01(Age)$$

In the third model, job satisfaction and total working years are statistically significant predictors, both with negative coefficients. Like Model 1, as job satisfaction increases by one or total working years increases by one, holding all else equal, the log-odds of attrition decreases by 0.22 and 0.05, respectively. Despite being statistically significant in Model 2, when adding additional predictors, age, along with the other employee demographic predictors, are not statistically significant. This indicates that while there may be a relationship between the employee demographics and attrition, it is not statistically significant and should not be used when assessing the likelihood of an employee leaving the company. *Refer to Appendix #2 for a summary of the first three models.*

The accuracy, balanced accuracy, area under the curve ("AUC"), AIC, and residual deviance of each model was assessed to determine the starting point for my final model. Accuracy is defined as the total correct predictions (of both attrition and no attrition) over the total predictions. Balanced accuracy is the calculated as the average of the sensitivity (the true positive rate) and specificity (the true negative rate) and was used given the imbalance in the original training and testing data. Area under the curve measures ranking of predictions and actuals. Finally, AIC and residual deviance are measures of model fit. *Refer to Appendix #3 for the models' metrics.*

While Model 1 displayed the best performance in four of the five metrics, given the same two predictors were statistically significant and the differences in metrics are generally immaterial (within 1-2%), I used Model 3 as the initial model when determining the top two predictors.

Final Model: Top Two Predictors

Using Model 3 as the starting point, I performed an "exhaustive" stepwise regression to determine the model of best fit. AIC was set as the metric to minimize when analyzing potential models and the under-sampled training set was used for model training. The model with the lowest AIC is predicted by job satisfaction and total working years, the same two statistically significant predictors in both Model 1 and Model 3. Once again, given the negative coefficients, as job satisfaction increases by one or total working years increases by one, holding all-else equal, the log-odds of attrition decreases by 0.23 and 0.07, respectively. The final model equation is as follows:

$$logit(p) = 1.13 - 0.23(JobSatisfaction) - 0.07(TotalWorkingYears)$$

The final model was used to predict on the training set (without sampling changes) and the testing set. The table below highlights the metrics from the training and testing sets.

	Training Set	Testing Set
Accuracy	0.5799	0.5745
Balanced Accuracy	0.6330	0.6359
AUC	0.6330	0.6360

The results of the testing set closely mirror that of the training set. Our final model predicted employee attrition and lack of attrition 57.45% of the time. *Refer to Appendix #4 for a summary of the final model.*

Recommendations

Given the analysis performed, it is evident that management's initial hypothesis was largely proven true; the top two predictors of attrition or lack of attrition are the employee's job satisfaction score and his/her total years of work experience. The employee demographics of gender, education, and age are not statistically significant predictors in determining attrition. As a result, I recommend a two-pronged approach to reduce the high levels of employee turnover:

1. Boost employee morale: Job satisfaction is a subjective metric as one person's definition of satisfaction may differ from another's definition. However, enterprise-wide initiatives should be implemented to improve employee morale. Employees already typically score job satisfaction as a three or four, indicating generally positive levels of satisfaction. However, additional initiatives such as spot bonuses for work ethic, real-time recognition to show employee appreciation, and employee events to boost camaraderie will improve employee morale and boost job satisfaction in employees. Given the subjectivity, the company should have employees' managers hold one-on-one meetings to gather feedback on what the employee wants and how they define job satisfaction so initiatives can be focused for each person.
2. Hire experienced employees: The current employee base has mostly worked less than 10 years and has worked at one additional company. Rather than focusing hiring initiatives on entry-level employees, the company should focus efforts on experienced hires. Using external headhunters may prove useful in finding candidates with more years of experience. In turn, this should reduce levels of future turnover.

While two potential solutions are proposed above, I recommend focusing initial efforts on retaining staff to prevent current burnout and turnover before focusing on external hires. Keeping talent within the organization and showing lower levels of turnover makes the company more attractive to experienced candidates looking for a job / company to settle with long-term. Aside from these two predictors, the company needs to ensure that controllable metrics such as total compensation remain competitive with other companies. Candidates then can overlook the compensation aspect and focus on the culture within the company (i.e., employees' job satisfaction).

Appendix

Appendix #1: Data Sampling

	No Sampling	Over and Under	Over	Under	ROSE (Synthetic Data)
Attrition	499	1,538	2,566	499	1,538
	(16.2%)	*(49.8%)*	*(49.8%)*	*(50.9%)*	*(49.8%)*
No Attrition	2,588	1,549	2,588	481	1,549
	(83.8%)	*(50.2%)*	*(50.2%)*	*(49.1%)*	*(50.2%)*

Appendix #2: Model 1, Model 2, and Model 3 Summary

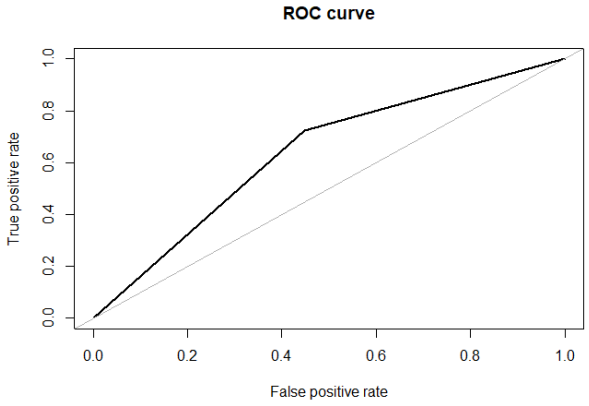
Regression Results

	Dependent variable:		
	Attrition		
	(1)	(2)	(3)
JobSatisfaction	-0.224*** (0.060)		-0.221*** (0.060)
TotalWorkingYears	-0.064*** (0.012)		-0.053*** (0.016)
YearsAtCompany	-0.007 (0.016)		-0.010 (0.016)
GenderMale		0.118 (0.134)	0.086 (0.136)
Education		-0.064 (0.063)	-0.055 (0.064)
Age		-0.040*** (0.007)	-0.011 (0.010)
Constant	1.307*** (0.201)	1.592*** (0.336)	1.700*** (0.394)
Observations	980	980	980
Log Likelihood	-643.602	-660.495	-642.452
Akaike Inf. Crit.	1,295.205	1,328.991	1,298.904
Note:			*p<0.1; **p<0.05; ***p<0.01

	Model 1 VIF	Model 2 VIF	Model 3 VIF
JobSatisfaction	1.014	N/A	1.015
TotalWorkingYears	1.781	N/A	2.809
YearsAtCompany	1.767	N/A	1.806
Gender	N/A	1.002	1.006
Education	N/A	1.004	1.006
Age	N/A	1.005	1.803

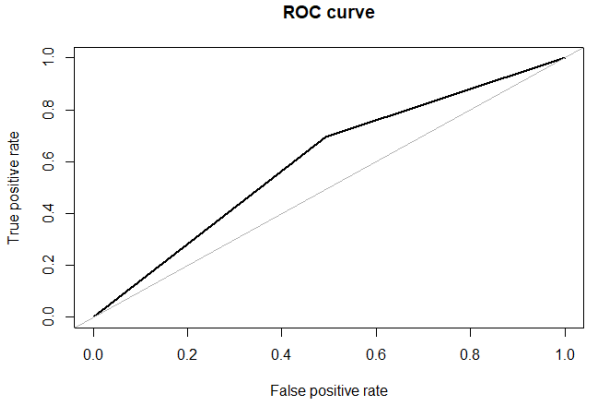
Appendix #2a: Model 1

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1428  138
         1 1160  361

               Accuracy : 0.5795
                 95% CI : (0.5619, 0.597)
    No Information Rate : 0.8384
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1507

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

            Sensitivity : 0.7234
            Specificity : 0.5518
         Pos Pred Value : 0.2373
         Neg Pred Value : 0.9119
             Prevalence : 0.1616
         Detection Rate : 0.1169
   Detection Prevalence : 0.4927
      Balanced Accuracy : 0.6376

       'Positive' Class : 1
```



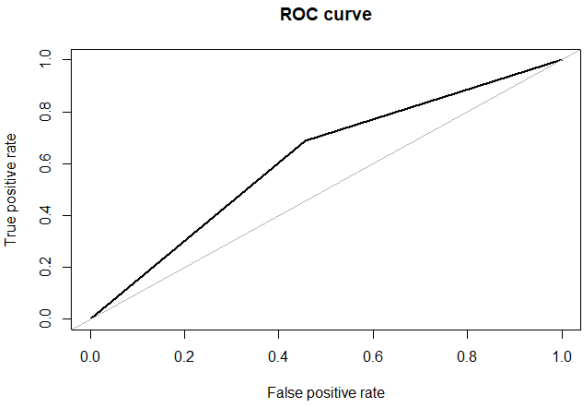
Appendix #2b: Model 2

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1313  152
         1 1275  347

               Accuracy : 0.5377
                 95% CI : (0.52, 0.5554)
    No Information Rate : 0.8384
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1062

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

            Sensitivity : 0.6954
            Specificity : 0.5073
         Pos Pred Value : 0.2139
         Neg Pred Value : 0.8962
             Prevalence : 0.1616
         Detection Rate : 0.1124
   Detection Prevalence : 0.5254
      Balanced Accuracy : 0.6014

       'Positive' Class : 1
```



Appendix #2c: Model 3

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1410  155
         1 1178  344

               Accuracy : 0.5682
                 95% CI : (0.5505, 0.5858)
    No Information Rate : 0.8384
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1282

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

            Sensitivity : 0.6894
            Specificity : 0.5448
         Pos Pred Value : 0.2260
         Neg Pred Value : 0.9010
             Prevalence : 0.1616
         Detection Rate : 0.1114
   Detection Prevalence : 0.4930
      Balanced Accuracy : 0.6171

       'Positive' Class : 1
```



Appendix #3: Model 1, Model 2, and Model 3 Metrics

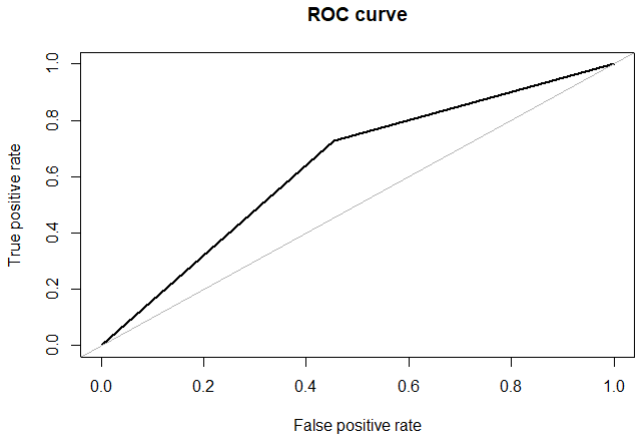
	Model 1	Model 2	Model 3
Accuracy	0.5795	0.5377	0.5682
Balanced Accuracy	0.6376	0.6014	0.6171
AUC	0.6380	0.6010	0.6170
AIC	1,295.2	1,329.0	1,298.9
Residual Deviance	1,287.2	1,321.0	1,284.9

Appendix #4: Final Summary

```
Call:
glm(formula = Attrition ~ JobSatisfaction + TotalWorkingYears,
    family = "binomial", data = employee_data_train_under)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6266  -1.1585   0.7871   1.0587   2.1847

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.306365   0.200617   6.512 7.43e-11 ***
JobSatisfaction  -0.225969   0.059935  -3.770 0.000163 ***
TotalWorkingYears -0.067311   0.009317  -7.224 5.04e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1358.2  on 979  degrees of freedom
Residual deviance: 1287.4  on 977  degrees of freedom
AIC: 1293.4

Number of Fisher Scoring iterations: 4
```

Our final model maximizes the sensitivity which, in other words, is a minimization of the false negative rate. False negatives are employees Canterra expects to stay but end up leaving. These present a higher risk than false positives (employees expected to leave but end up staying), as they bring a level of surprise that results in the detriment to the company. Thus, the minimization of false negative rate (which is 27.4%) is desired.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 606  58
         1 505 154

               Accuracy : 0.5745
                 95% CI : (0.5473, 0.6013)
    No Information Rate : 0.8398
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1467

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

            Sensitivity : 0.7264
            Specificity : 0.5455
         Pos Pred Value : 0.2337
         Neg Pred Value : 0.9127
             Prevalence : 0.1602
         Detection Rate : 0.1164
   Detection Prevalence : 0.4981
      Balanced Accuracy : 0.6359

       'Positive' Class : 1
```



	Final Model VIF
JobSatisfaction	1.008
TotalWorkingYears	1.008