CANTERRA lOGISTIC rEGRESSION

Abigail Hoffman

Georgetown University

Machine Learning I

1. **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 that 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 company and a longer tenure of total working years reduces 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 if an employee left the company or not (attrition).

1. **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/no the employee did not leave), a logistic regression is determined to be the best model type to understand the data. Senior management’s hypothesis that 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 as 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. [Insert Recommendation]

1. **Data Exploration**

The data set provided by Canterra offers data for 18 variables (columns) and 4410 employees at the company 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.

1. **Methods & Models**

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

1. 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]](#footnote-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 and for purposes of AIC, lower is better.

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

1. Modeling

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

*Logistic Regression (Log-Odds)*

The final model’s significant variables were selected by the stepwise method.[[2]](#footnote-2) The five variables in the regression model below were determined to be a predictor of attrition.

See appendix part B for model (5) results.

1. **Results**

**Table

Description automatically generated**Based on the final model, management’s hypothesis was correct that a higher survey result of job satisfaction does reduce 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 to 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, I added in number of training times last year to see if the opportunity to engage with work colleagues and new or review material and topics. This did result in 13.3 percent decrease in attrition with just one additional training. This is an area of opportunity to increase opportunities for trainings, quarterly for example, the odds of attrition would decrease four times (53.2 percent). Age and total working variables were significant for the model; however, the beta percentage is marginal.

1. **Model Performance**

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

**Text

Description automatically generated**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 sensitivity or the power of the model to detect attrition cases is 68 percent when the threshold is 0.5, or 50 percent. Other notable factors include the prevalence rate of the amount of attrition occurring in the training set at 15.91 percent and the model’s detection rate is 10.82 percent. The models AUC, area under the curve, is 0.698. Our model’s performance is above average and ready to be tested on the actual test data set.

A screenshot of a computer

Description automatically generated with low confidenceAlthough 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, or 50 percent. Other notable factors include the prevalence rate of 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 models AUC, area under the curve, is 0.672. There is a 67.2 percent chance that the model can differentiate between positive class and 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.

**Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated**The model’s performance did slightly decrease in efficiency; however, the test set had a larger value of observations, 1,323. To best fit the training set, the under sampling as described in methods and models’ section was selected to best fit the logistic model for performance. 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 at the attrition y values of varying levels.

1. **Recommendations**

Canterra has a multitude of opportunities for improvement to reduce attrition and increase employee satisfaction. After statistical analysis, job satisfaction was the largest indicator of reduced attrition. The happier the employee, the less likely they are to leave. One way to possibility is to use a network like Pulse Surveys on a quarterly basis to understand employee well-being, departmental pain points and understand 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, potentially allowing hybrid workdays with flex schedules that empowers workers to set their own days and schedules could positively influence attrition. 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 workers but find other people who may be able to help or support them when a difficult task arises. Another idea is offering department shadowing, where someone may be more interested in another department’s work and can spend a week with that department to see if that is the right fit. This would reduce turnover as employees stay within the company rather than seeking external opportunities.

**Appendix**

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

Table

Description automatically generated

1. Github Link to all materials for reproducibility

1. R Documentation [↑](#footnote-ref-1)
2. Stepwise regression (bidirectional) is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure in R studio. [↑](#footnote-ref-2)
3. Canterra Employee Data [↑](#footnote-ref-3)