Ashley Daniels and Andrey McLeggan
OPAN 6602 - ML I
Group 7
Project 2 Written Report

Canterra, a large organization with approximately 4,000 employees, faces an annual attrition rate of 15%. This high level of turnover poses significant challenges, including project delays that impact timelines and reputation, increased costs associated with maintaining a dedicated recruitment department, and additional expenses for training and acclimatizing new employees. Concerned about these issues, the management has hypothesized that factors like job satisfaction and total working years may influence attrition rates. Furthermore, they are interested in exploring the impact of demographic variables such as age, gender, and education on employee retention.

To address these questions, Canterra seeks to identify actionable strategies to reduce attrition by analyzing internal and external recruitment processes. Logistic regression emerges as a more appropriate analytical tool than linear regression for this task, given the binary nature of attrition (i.e., whether an employee leaves or stays). By employing logistic regression, we can uncover the likelihood of attrition based on the identified factors, providing more reliable and interpretable insights for decision-making.

Logistic regression is the best modeling technique for analyzing this dataset due to the dependent variable of attrition being 1's and 0's (binary). The logistic regression is specifically made for this type of outcome vs the classical regression that favors continuous variables. This allows the predictions to be bounded to numbers between 0-1, such as .074, which means that 74% of attrition is based on x. In linear regression, the numbers can be higher than or lower than 0-1, not allowing for easier interpretation. The most significant reason why logistical is more appropriate is classic regression assumes a linear relationship with which the nature of our dependent variable is 0-1, these binary outcomes often exhibit a nonlinear relationship which is handled better by logistic regression. Additionally, Logistic regression can predict the probability of attrition for employees of different ages. For example, you might find that employees aged 25 have a 30% probability of leaving, while those aged 55 have a 5% probability, enabling targeted interventions.

The best-fit model was determined to be model four, which evaluated interactions of attrition involving age, gender and job satisfaction. The model was a multiple logistic regression model analyzing the binary of yes (1) or no (0) employee attrition. The summary of this model provided key insights into the statistical significance and impact of these predictors on the likelihood of attrition. For example, This analysis is an ideal basis for Canterra to proactively identify employees who are most likely to leave the company. In turn, this allows the company to strategize how to create more favorable circumstances to deter prospective leavers from leaving the company.

Marginal distribution is the distribution of a variable within the dataset ignoring relationships or dependencies between the variable and other variables, showing the frequency of the variable across the entire dataset. The distribution in this case shows the probability of attrition relative to the total number of employee population. In this case, we see the company has a 16% attrition rate and 84% retention rate.

Attrition	Percentage
No	83.88287
Yes	16.11713

Evaluating the marginal distribution of the employees who stayed with Canterra and those who left, showed that most employees remained in comparison to those who left. Canterra has a less than 20% attrition rate which is low compared to the number of employees who have decided to stay. It is also evident that younger employees (=>40 years old) are more likely than older employees to leave the company. Younger employees are about 18.37% more likely to leave Canterra than their 41+ counterparts at 11.26%.

Age	Attrition Percentage	
Less than or equal to 40yrs	18.37%	
Greater than 40yrs	11.26%	

A logistic regression model illustrating the relationship between age and attrition is located in Figure X. The plot includes pink dots representing actual data points of age and attrition. The zeroes and ones indicate employees who stayed and those who left. The green line represents the predicted probability of attrition based on age. According to the model, younger employees are more likely to leave the company, as they have a higher probability of attrition, while older employees are less likely to leave, showing a lower probability of attrition. A summary table, labelled ‘age_summary’, to compare the ages of employees who stayed at Canterra versus those who left was also created. This table supports our model’s findings, i.e. indicating that younger employees are more likely to leave. Specifically, the table shows that employees who stayed tend to be older on average with a smaller variation in their age range. For example, the median age of employees who stayed is 36 years, while the median age of those who left is 31 years. Additionally, the lowest age in both groups is 18 years old, but the oldest employee in the staying group is 60, compared to 58 for those who left.

After running a logistic regression model for 4:

1. Model 1: A one-variable model with `Age`.
2. Model 2: A two-variable model with `Age` and `Gender`
3. Model 3: A three-variable model with `Age`, `Gender`, and `JobSatisfaction`.
4. Model 4: An interaction model involving `Age`, `Gender`, `JobSatisfaction`, `Income`, and `Gender:Income`.

An analysis evaluating the validation data AIC, AUC, Precision, and Recall, as shown below, was created:

1. AIC: This shows the measure of model fit, the lower AIC indicates better trade-off between best fit and complexity.
2. AUC: Measures the model's ability to see the differences in class, a higher AUC shows better performance with 1 being ideal, .5 being random, and anything lower being not good.
3. Precision: Calculated with the formula TP/TP+FP, shows the portion of predicted positives that were actually positive; emphasizes avoiding FP(false positives). The higher the better 0-1.
4. Recall: Calculated by TP/TP+FN, shows the portion of actual positive cases that the model correctly depicted. The higher the better 0-1.

Model Comparison Grid

Model	AIC	AUC	Precision	Recall
Model_1	3453.735	.643	.596	.693
Model_2	3453.342	.628	.583	.66
Model_3	3430.298	.642	.594	.62
Mode_4	3414.859	.64	.602	.643

AIC (Akaike Information Criterion):

- A lower AIC indicates a better fit of the model to the data. Model 3 has the lowest AIC (3430.298), making it the best in terms of model fit.

AUC (Area Under the Curve):

- A higher AUC indicates better overall classification performance. Model 1 has the highest AUC (0.643), followed closely by Model 3 (0.642).

Precision:

- Precision measures how many of the predicted positives are true positives. Model 4 has the highest Precision (0.602), indicating it is best at correctly identifying true positives.

Recall:

- Recall measures how many of the actual positives are correctly predicted. Model 1 has the highest Recall (0.693), meaning it performs best at capturing true positives.

Overall, Model 3 has the lowest AIC and strong AUC, Precision, and Recall, making it the most balanced choice overall. The data in the chart above were achieved after rebalancing our data with oversampling to ensure good representation on both sides of the "yes" and "no" then evaluate for the AUC, AIC, precision as well as recall using the validation dataset.

While our best-performing model showed some promise, it still has room for improvement. The AUC value of this model used on the test set was higher than 0.5, indicating that it outperforms random guessing, but it is still far from ideal. In this analysis, we tested four logistic regression models to predict employee attrition, each using different sets of predictor variables such as Age, Gender, Job Satisfaction, and Income. The fourth model, which used only Age, Gender, Job Satisfaction, Income and the relationships between Gender and Income as predictors, achieved the highest AUC of 0.61, making it the most effective at distinguishing between employees who left and those who stayed.

To assess the discriminatory power of each model, we plotted the ROC curves for all four models on a single graph. While model four performed the best, its AUC still suggested limited predictive strength. The results highlight the need for further refinement, such as incorporating additional relevant predictors or exploring more advanced modeling techniques, to improve the accuracy of attrition predictions.

Our analysis revealed that the attrition rate at Canterra is strongly influenced by age and job satisfaction. Slightly over 16% of the total employees in the dataset provided decided to leave the company as shown in the table . As previously mentioned, the median age of employees who stayed with the company was 36 years old, meanwhile the median age of leavers was 31 years old. The difference in median age highlights that age is an important factor to predict future attrition. The exploratory data analysis (EDA) showed that job satisfaction, gender, and income can also help indicate which populations of employees are more likely to leave or stay.

Utilizing logistic regression modeling comparisons, we were able to demonstrate that as average employee age increases probability of leaving the company generally decreases which

aligns with the EDA insights. Additionally, using the ROC curve as another method of analysis helped to confirm the reliability of the final model.

As age seems to be a huge driving factor in attrition, it is recommended that Canterra do a focused target effort on younger employees. This can be done by establishing a mentorship program with senior employees at the company and establishing programs such as leadership development programs. This program can allow younger employees to grow fast in the company and gain exposure to many lines of business while allowing them to be guided and grown by seniors. This can increase retention as these employees can grow within the company and see that there's a bright future ahead at canterra with career flexibility. This can be a strong initiative especially when paired with data from employee NPS surveys. This can capture pain points amongst employees and gauge where the dissatisfaction is stemming from even more clearly. It is believed that doing these things along with rewarding high performers and tenured individuals with discretionary bonus will show that canterra to showcase that they indeed not only care for the future of their young employees but care about their performance and tenure. This can dramatically improve culture which overtime, can be evaluated with the change in sentiments over time by analyzing the NPS survey data.

Appendix 1 – Analysis Visuals and Tables

Figure 1 – Age v. Probability of Employee Attrition

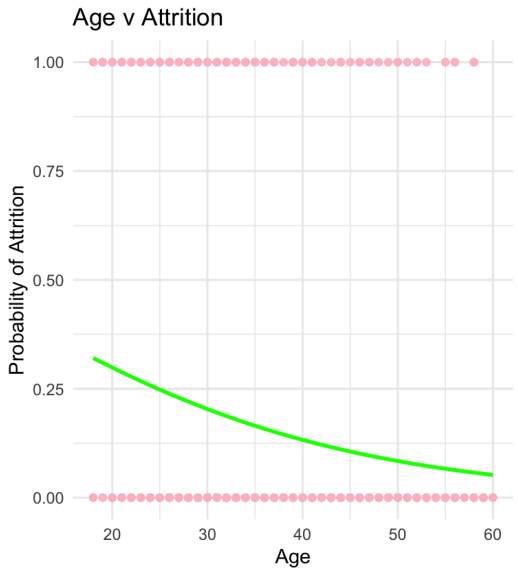
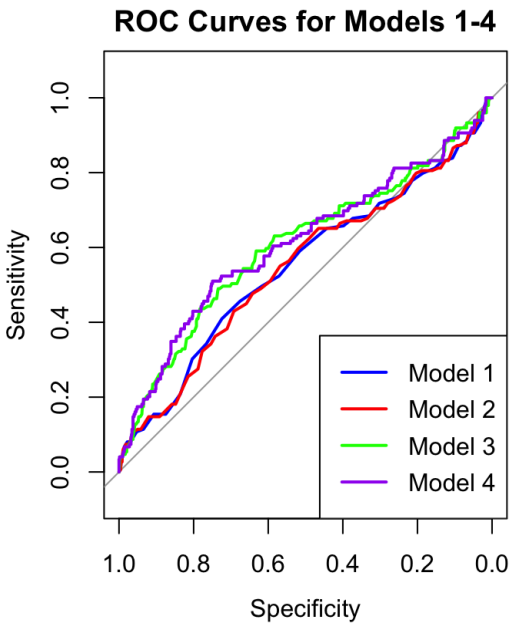


Figure 2 – Age v. Probability of Attrition Summary Statistics

Attrition	Count	Mean Age	SD Age	Min Age	Max Age	Median Age
Yes	699	33.672	9.697	18	58	32
No	3638	37.573	8.91	18	60	36

Figure 3 – ROC Model on Test Set


ROC Curves for Models 1-4
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Model 1
Model 2
Model 3
Model 4