**+Background**

A large company named Canterra, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons:

1. The former employees’ projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners
2. A sizable department has to be maintained, for the purposes of recruiting new talent
3. More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company

The management hypothesizes that higher job satisfaction and a higher number of total working years will reduce employee attrition. Additionally, the marketing management was interested to know if demographic variables such as gender, education and age affect employee attrition. Hence, the management has contracted you as a consultant to understand whether these two factors they should focus on, in order to curb attrition. In other words, they want to know if changes in their internal and external recruitment strategies would help retain employees.

There are 18 variables (columns) and 4410 employees as described on the [Canterra Employee Dataset – Projects](https://georgetown.instructure.com/courses/197546/pages/canterra-employee-dataset-projects-2-individual-and-3-team) page in the Getting Started section.

**Project 2 Goal and Key Questions**

**Goal**You are required to model the probability of attrition and draw insights from your analysis.

**Key Questions**

1. Logistic Regression Rationale
   1. ~~Explain why logistic regression is an appropriate modeling technique for predicting employee attrition in this dataset compared to classical regression methods.~~

* ~~Show the comparison between linear regression and logistic regression in appendix.~~

1. ~~Provide a specific rationale for utilizing logistic regression in predicting attrition concerning the age of employees.~~
2. ~~Model Parameter Interpretation~~
   1. ~~Fit a logistic regression model~~
   2. ~~Summary of the model~~
   3. ~~After conducting the logistic regression analysis, interpret the parameter estimates obtained from the model with a focus on their significance in predicting employee attrition.~~
3. Distribution Estimates
   1. ~~Describe the concept of a marginal distribution and explain its relevance in understanding employee attrition.  Calculate and report the estimated marginal distribution of attrition (Attrition: Yes/No) without the influence of the logistic regression model.~~
   2. ~~For modeling purposes, choose an age value representing a "younger" employee and another age value for an "older" employee. Calculate and report the model-based estimates of the conditional distribution of attrition for these two age groups.~~
4. Graphical Representation
   1. ~~Create a graph illustrating the relationship between age and attrition using the logistic regression model. Include appropriate labels and explanations. (10 points)~~
   2. ~~Create a table summarizing the key statistics of the age variable in relation to attrition and discuss any insights from the table.~~
5. Model Comparison - Choose four variables from the list provided (e.g., `Age`, `Gender`, `JobSatisfaction`, `Income`) and use them to create logistic regression models with the following combinations:
   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`.
   5. ~~Compare the validation-set performance of these models using the following metrics: AIC, AUC, Precision & Recall (use p = 0.5 as cutoff).~~ Include tables summarizing the results.
   6. ~~Create a graph to visually represent the comparison of the models based on the selected metrics.~~
6. ~~Predictive Performance - Using the logistic regression models that you have developed, perform a prediction analysis for employee attrition.~~
   1. ~~For each of the four models, calculate the predicted probabilities of employee attrition for the test dataset.~~
      1. ~~Calculate predicted probabilities for each model using the test dataset~~
      2. ~~Create ROC curve objects for each model~~
   2. ~~Create ROC curves for each model to evaluate their discriminatory power in predicting attrition.~~
      1. ~~Plot ROC curves for all four models on one plot~~
   3. ~~Compare the ROC curves and AUC. Discuss which model appears to be the most effective in distinguishing between employees with and without attrition.~~
7. Recommendations and Reflection
   1. ~~Summarize the key findings from the analysis, including insights gained from the EDA, model building, and model comparison. Include tables or visualizations to support your summary.~~
   2. ~~Provide recommendations to Canterra based on the analysis. Suggest strategies to reduce attrition rates and improve employee retention.~~

**Deliverable**

Submit a professional report in a PDF document. Include direct explanations for each question in an appendix. Include your code in another appendix.

Your formal project report should clearly answer the questions and show the results of the analysis as requested. We will grade your report on the clarity of the answers and comprehension, depth of analysis, and completeness of the answers.

To summarize:

* Submit a professional report in pdf.
* You must support your answers with evidence.
* Your report should be carefully organized, clearly and concisely written, and must answer all questions.

Appendix A. Model Summary Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AIC | AUC | Residual Deviance | Fisher Scoring |
| 1 | 811.31 |  |  |  |
| 3 | 689.66 | 0.7736 | 631.66 | 5 |
| 4 | 891.36 | 0.6265 | 887.36 | 4 |
| 5 | 730.32 | 0.6265 | 726.32 | 4 |
| 6 | 731.62 | 0.6262 | 725.62 | 4 |
| 7 | 722.05 | 0.6519 | 708.05 | 4 |
| 8 | 723.83 | 0.6517 | 705.83 | 4 |

Appendix B. Linear Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | t.value | Pr |
| (Intercept) | 1.35982203 | 0.28844747 | 4.71427968 | 2.9863E-06 |
| Age | -0.0368462 | 0.01386225 | -2.6580242 | 0.00805888 |
| BusinessTravelTravel\_Frequently | 0.25271938 | 0.08517808 | 2.9669531 | 0.00312131 |
| BusinessTravelTravel\_Rarely | 0.03201568 | 0.07914057 | 0.40454196 | 0.6859511 |
| MaritalStatusMarried | 0.05386615 | 0.05223259 | 1.03127468 | 0.30280651 |
| MaritalStatusSingle | 0.20720024 | 0.05320497 | 3.89437717 | 0.00010893 |
| Income | 3.6193E-07 | 1.5605E-06 | 0.23193075 | 0.81666688 |
| NumCompaniesWorked1 | 0.11297744 | 0.05881625 | 1.92085435 | 0.05519965 |
| NumCompaniesWorked3 | 0.04122093 | 0.08151675 | 0.50567436 | 0.61326159 |
| NumCompaniesWorked4 | 0.08119061 | 0.08026919 | 1.01147912 | 0.31217425 |
| NumCompaniesWorked5 | 0.30243074 | 0.0942569 | 3.20857939 | 0.00140139 |
| NumCompaniesWorked6 | 0.19014847 | 0.09511679 | 1.99910527 | 0.04602469 |
| NumCompaniesWorked7 | 0.29667662 | 0.09011569 | 3.29217514 | 0.00104955 |
| NumCompaniesWorked9 | 0.24997547 | 0.10517776 | 2.37669518 | 0.01776534 |
| TrainingTimesLastYear | -0.03263973 | 0.01476704 | -2.21030917 | 0.02744122 |
| YearsWithCurrManager | -0.02348202 | 0.00590604 | -3.97593207 | 7.8203E-05 |
| EnvironmentSatisfaction2 | -0.11228896 | 0.05452235 | -2.05950312 | 0.03985511 |
| EnvironmentSatisfaction3 | -0.11065891 | 0.04990652 | -2.21732354 | 0.02695598 |
| EnvironmentSatisfaction4 | -0.10805617 | 0.04945297 | -2.18502904 | 0.02925322 |
| EnvironmentSatisfactionNA | 0.08544574 | 0.23284721 | 0.36696053 | 0.71377129 |
| JobSatisfaction2 | -0.16329186 | 0.05436578 | -3.00357801 | 0.00277375 |
| JobSatisfaction4 | -0.18779315 | 0.05248735 | -3.57787431 | 0.00037291 |
| Age\_squared | 0.00042088 | 0.00017605 | 2.39070232 | 0.01710788 |
| Income\_squared | -2.4591E-12 | 7.7063E-12 | -0.31910069 | 0.7497556 |

Appendix C. Model 1: Logistic Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | z.value | Pr |
| (Intercept) | 2.7666482 | 1.25468486 | 2.20505426 | 0.02745029 |
| Age | -0.01277243 | 0.01544117 | -0.82716746 | 0.40814215 |
| BusinessTravelTravel\_Frequently | 1.37694156 | 0.46698373 | 2.9485857 | 0.00319232 |
| BusinessTravelTravel\_Rarely | 0.2624456 | 0.42558792 | 0.61666599 | 0.53745506 |
| DistanceFromHome | -0.00620183 | 0.01285806 | -0.48233031 | 0.62957132 |
| GenderMale | -0.33905976 | 0.2146097 | -1.5798902 | 0.11413201 |
| JobLevel | -0.1854774 | 0.09580275 | -1.93603406 | 0.05286353 |
| MaritalStatusMarried | 0.36120814 | 0.28745477 | 1.25657379 | 0.20890801 |
| MaritalStatusSingle | 1.29597018 | 0.30183509 | 4.29363655 | 1.7577E-05 |
| NumCompaniesWorked4 | 0.50191807 | 0.4639454 | 1.08184727 | 0.2793204 |
| NumCompaniesWorked5 | 2.31067628 | 0.58963847 | 3.9188018 | 8.899E-05 |
| NumCompaniesWorked6 | 1.30766421 | 0.5814388 | 2.24901435 | 0.02451158 |
| NumCompaniesWorked7 | 1.84281095 | 0.58721266 | 3.13823435 | 0.00169969 |
| NumCompaniesWorked8 | 0.81133683 | 0.66047402 | 1.22841596 | 0.21929086 |
| NumCompaniesWorked9 | 1.33548105 | 0.59166254 | 2.25716683 | 0.02399765 |
| TrainingTimesLastYear | -0.19844403 | 0.08398031 | -2.36298274 | 0.01812851 |
| YearsWithCurrManager | -0.11120127 | 0.06064188 | -1.83373725 | 0.06669301 |
| EnvironmentSatisfaction2 | -0.75301503 | 0.30495858 | -2.46923706 | 0.01354015 |
| EnvironmentSatisfaction3 | -0.65557831 | 0.28497376 | -2.30048655 | 0.02142067 |
| EnvironmentSatisfaction4 | -0.72418798 | 0.28276818 | -2.56106606 | 0.01043515 |
| JobSatisfaction2 | -1.06763637 | 0.31935603 | -3.34309137 | 0.00082851 |
| JobSatisfaction3 | -0.85313913 | 0.29347686 | -2.90700643 | 0.00364906 |
| JobSatisfaction4 | -1.06683832 | 0.31131937 | -3.42682923 | 0.00061067 |

Appendix D. Model 4: Logistic Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | z.value | Pr |
| (Intercept) | 1.61233959 | 0.31227839 | 5.16314822 | 2.4283E-07 |
| Age | -0.04023763 | 0.00850183 | -4.73282025 | 2.2142E-06 |

Appendix E. Model 5: Logistic Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | z.value | Pr |
| (Intercept) | 0.38668741 | 0.38000229 | 1.01759231 | 0.30887175 |
| Age | -0.04159647 | 0.01057134 | -3.9348361 | 8.3253E-05 |

Appendix F. Model 6: Logistic Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | z.value | Pr |
| (Intercept) | 0.51551734 | 0.41024178 | 1.25661832 | 0.20889188 |
| Age | -0.0425276 | 0.01063715 | -3.99802656 | 6.3873E-05 |
| GenderMale | -0.1568651 | 0.18740365 | -0.83704399 | 0.40256784 |

Appendix G. Model 7: Logistic Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | z.value | Pr |
| (Intercept) | 1.09054731 | 0.45268409 | 2.40906923 | 0.01599326 |
| Age | -0.04405735 | 0.01079951 | -4.07957023 | 4.5119E-05 |
| GenderMale | -0.10515855 | 0.19040344 | -0.55229335 | 0.58074739 |
| JobSatisfaction2 | -0.8024758 | 0.27843338 | -2.88211065 | 0.00395021 |
| JobSatisfaction3 | -0.39770312 | 0.2452553 | -1.62158826 | 0.10489154 |
| JobSatisfaction4 | -1.04799061 | 0.27350029 | -3.83177153 | 0.00012722 |
| JobSatisfactionNA | -0.39584304 | 1.24352504 | -0.31832334 | 0.75023968 |

Appendix H. Model 3: Logistic Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | z.value | Pr |
| (Intercept) | 5.83507884 | 1.62261767 | 3.59608978 | 0.00032304 |
| Age | -0.31373532 | 0.07867687 | -3.98764383 | 6.6733E-05 |
| BusinessTravelTravel\_Frequently | 1.5026287 | 0.45493637 | 3.30294257 | 0.00095676 |
| BusinessTravelTravel\_Rarely | 0.57088022 | 0.42268333 | 1.35060972 | 0.17682049 |
| MaritalStatusMarried | -0.04303614 | 0.29348173 | -0.14663993 | 0.88341623 |
| MaritalStatusSingle | 0.88807125 | 0.2906022 | 3.0559688 | 0.00224335 |
| Income | 3.6014E-06 | 8.6143E-06 | 0.41806647 | 0.67589852 |
| NumCompaniesWorked1 | 0.344208 | 0.32819745 | 1.04878328 | 0.29427787 |
| NumCompaniesWorked2 | -0.58993846 | 0.48553036 | -1.21503926 | 0.22435112 |
| NumCompaniesWorked3 | -0.02026242 | 0.47559907 | -0.042604 | 0.96601721 |
| NumCompaniesWorked4 | 0.37493938 | 0.46416701 | 0.80776825 | 0.41922401 |
| NumCompaniesWorked5 | 1.00902766 | 0.57929311 | 1.74182576 | 0.08153894 |
| NumCompaniesWorked6 | 0.80515544 | 0.51192577 | 1.57279726 | 0.11576577 |
| NumCompaniesWorked7 | 1.30876621 | 0.50717569 | 2.58049868 | 0.00986577 |
| NumCompaniesWorked8 | 0.62968913 | 0.60004131 | 1.04940963 | 0.29398963 |
| NumCompaniesWorked9 | 1.1948971 | 0.54198775 | 2.20465705 | 0.02747817 |
| YearsWithCurrManager | -0.09767673 | 0.03280686 | -2.97732612 | 0.00290775 |
| EnvironmentSatisfaction2 | -0.61646407 | 0.30507764 | -2.02067931 | 0.04331297 |
| EnvironmentSatisfaction3 | -0.72247155 | 0.28800355 | -2.50855084 | 0.01212275 |
| EnvironmentSatisfaction4 | -0.65312707 | 0.27714996 | -2.35658371 | 0.01844391 |
| JobSatisfaction2 | -0.90075858 | 0.30316805 | -2.97115272 | 0.00296684 |
| JobSatisfaction3 | -0.66723162 | 0.27182027 | -2.45467937 | 0.01410103 |
| JobSatisfaction4 | -1.22020014 | 0.29508698 | -4.13505251 | 3.5487E-05 |
| Age\_squared | 0.00364786 | 0.00098868 | 3.68964762 | 0.00022456 |
| Income\_squared | -1.9574E-11 | 4.3109E-11 | -0.45406422 | 0.64978261 |

Appendix I. Model 8: Logistic Regression Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Estimate | Std.Error | z.value | Pr |
| (Intercept) | 1.4250598 | 0.50879319 | 2.80086256 | 0.00509662 |
| Age | -0.04426894 | 0.01082479 | -4.08959 | 4.3214E-05 |
| factor(Gender)Male | -0.43046721 | 0.33024482 | -1.30347908 | 0.19241126 |
| JobSatisfaction2 | -0.83161646 | 0.27946141 | -2.97578277 | 0.00292242 |
| JobSatisfaction3 | -0.41234641 | 0.24626288 | -1.6744156 | 0.09404895 |
| JobSatisfaction4 | -1.04551599 | 0.27405689 | -3.81495964 | 0.00013621 |
| JobSatisfactionNA | -0.45841486 | 1.24618119 | -0.3678557 | 0.71298083 |
| Income | -4.8075E-06 | 3.3282E-06 | -1.44448229 | 0.14860336 |
| factor(Gender)Male:Income | 4.9351E-06 | 4.1894E-06 | 1.17799513 | 0.23879855 |

Appendix J. Age Variable Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attrition | Mean | Median | Standard Deviation | Minimum Age | Max Age |
| No | 37.5 | 36 | 9.02 | 18 | 60 |
| Yes | 33.7 | 32 | 9.76 | 18 | 58 |

Appendix K. Sample of Predicted Probability for Test Data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Income | Marital Status | Business Travel | Model3 | Model5 | Model6 | Model7 | Model8 |
| 51 | 131160 | Married | Travel\_Rarely | 0.05744095 | 0.1499832 | 0.16065128 | 0.0993589 | 0.07524894 |
| 38 | 83210 | Married | Non-Travel | 0.01682129 | 0.23254811 | 0.22142239 | 0.14972224 | 0.15154651 |
| 55 | 55380 | Single | Travel\_Rarely | 0.24874593 | 0.12998191 | 0.13901436 | 0.20872845 | 0.21823686 |
| 28 | 25920 | Married | Travel\_Rarely | 0.14227255 | 0.3147481 | 0.30319688 | 0.21480394 | 0.21633637 |
| 21 | 42130 | Single | Travel\_Rarely | 0.703156 | 0.3806386 | 0.36948487 | 0.41641233 | 0.41531782 |
| 26 | 104470 | Divorced | Travel\_Frequently | 0.38856738 | 0.33296149 | 0.32146108 | 0.27636229 | 0.27394818 |
| 53 | 21480 | Married | Travel\_Rarely | 0.11425029 | 0.13968295 | 0.1495105 | 0.16216655 | 0.19204055 |
| 42 | 89260 | Married | Travel\_Rarely | 0.07381471 | 0.20418106 | 0.19348743 | 0.22050547 | 0.22000166 |
| 29 | 65130 | Single | Travel\_Frequently | 0.39654415 | 0.30584638 | 0.29428804 | 0.20746647 | 0.20975384 |
| 26 | 162910 | Married | Travel\_Frequently | 0.75567951 | 0.33296149 | 0.35658891 | 0.48626749 | 0.37539955 |
| 26 | 68540 | Divorced | Travel\_Rarely | 0.14132871 | 0.33296149 | 0.32146108 | 0.27636229 | 0.273037 |
| 28 | 96370 | Single | Travel\_Rarely | 0.20175781 | 0.3147481 | 0.30319688 | 0.21480394 | 0.21786468 |
| 26 | 157870 | Single | Travel\_Rarely | 0.22720102 | 0.33296149 | 0.32146108 | 0.23003864 | 0.23473476 |
| 18 | 38120 | Single | Travel\_Rarely | 0.79715962 | 0.41046362 | 0.39966898 | 0.44884296 | 0.44775902 |
| 32 | 27060 | Married | Travel\_Rarely | 0.0827177 | 0.28001246 | 0.26850302 | 0.30530754 | 0.30345351 |
| 31 | 132250 | Married | Travel\_Rarely | 0.46536008 | 0.28847476 | 0.2769376 | 0.40603304 | 0.4107502 |
| 39 | 196650 | Married | Travel\_Rarely | 0.05148307 | 0.22520714 | 0.21417779 | 0.32457007 | 0.330307 |