Canterra Employee Churn

Determining Significant Factors to Curb Attrition

Saxa - Team 9: Kassandra Sellers, Clark Necciai, Mabel Barba, Benny Huang, Tony Campoverde

2023-12-11

Contents

1 Executive Summary

We at People Analytics Consulting Firm had been previously approached by Canterra in helping to assess the most important factors relating to their high employee attrition. Within their organization, approximately 15% of employees leave the company every year, leading management to speculate on the causes of such high attrition rates. In order to gain a deeper level of understanding as to the primary factors Canterra should focus on in curbing attrition, we again evaluated individual employee demographics at a granular level.

Our team first performed multivariate analysis and deliberated on key demographics to consider in the modeling process. Resulting models such as logistic regression, decision trees, and random forests balanced explainability and accuracy in determining employee attrition from Canterra. The optimal random forest models provided the most insightful findings, revealing key demographics such as income, age, and the total amount of working years as being paramount in determining employee attrition. Canterra management can act on these findings to identify and engage with employees likely to experience attrition.

2 Problem Statement and Approach

Our primary tasks in this analysis were twofold:

- Determine influential variables through Logistic Regression, Decision Tree, and Random Forest models
- Evaluate which model would best represent and explain findings to Canterra

We began with an introductory analysis, in which we evaluated univariate and multivariate distributions and their relationships to the target variable. After our analysis of 18 features over 4410 observations, we generated Logistic, Decision Tree, and Random Forest models to further extrapolate employees' most telling features in determining their possible attrition from Canterra. After thorough deliberation as to which model would best represent our findings, our team agreed that presenting our findings from the perspective of the Random Forest model would be the optimal approach. Due to the reliable, robust performance of the random forest model in predicting employee attrition, we have confidence in our resulting recommendations.

3 Methodology

3.1 Data Preprocessing

Our team began with an inspection of Canterra's provided employee dataset. We discovered no duplicate observations, but did identify few instances of missing values which were promptly removed due to their random placement and negligible quantity[Fig#1]. Variables `JobLevel`, `Education`, `NumCompaniesWorked`, and `TotalWorkingYears` were determined to have non-representative data types of which were promptly reevaluated to be more representative of their underlying values[Fig#2]. Additionally, in an effort to maintain employee confidentially and remove irrelevant predictors in the analytical process, we opted to remove the `EmployeeID` variable. At times, our coefficient estimates become incredibly small when we have continuous variables with large scales. To better represent the effects that `Income` may have in determining `Attrition`, we opted to scale the `Income` variable, as doing so does not change the interpretability and provides for a stable `Income` coefficient.

3.2 Introductory Analysis

3.2.1 Target Variable

3.2.1.1 `Attrition` Canterra had previously reported that approximately 15% of their employees leave the company every year. This assertion is in agreement with our evaluation of the `Attrition` distribution[Fig#3]. According to the marginal distribution of `Attrition`, we see that approximately 16.1% of our observations are of employees which have experienced attrition. A significant majority of our target variable observations are employees which did not experience attrition. Because we see so many employees that did not experience attrition, our models will more than likely perform well in correctly predicting employees which did not

experience attrition, but perform poorly in correctly predicting employees which did experience attrition. Because of this, the resulting accuracy and specificity metrics describing model performance may be misleading as they do not take into account the distribution of classes.

Other metrics then become preferable, such as precision, recall/sensitivity, or the F1 Score. While we will still report the accuracy of our models' performance, because of our significant imbalance, we will consider other aspects of model performance in addition to accuracy.

3.2.2 Predictor Variables

3.2.2.1 Categorical/Factor Variables Our categorical variables include `BusinessTravel`, `Education`, `Gender`, `JobLevel`, `MaritalStatus`, `EnvironmentSatisfaction`, and `JobSatisfaction`. On the whole, none of these variables distributions are equally distributed$_{[Fig\#4/\#5]}$. Below we mention noteworthy findings:

- `BusinessTravel`: Nearly 70% of all values are of those employees which rarely travel for work
- `Education`: A majority of employees(38.9%) have attained Bachelor degrees. A mere 3% of employees have Doctorate degrees
- `Gender`: A slight imbalance between male and females exist, with nearly 60% of all Canterra employees being male
- `JobLevel`: A vast majority of employees occupy lower level positions, with 73.18% of the employees occupying job levels 1 and 2
- `MaritalStatus`: Most employees are currently married, followed secondly by single and lastly divorced. There appears to be a greater number of single employees which experience attrition
- `EnviromentSatisfaction/JobSatisfaction`: Most employees($\approx 60\%$) state that their environment and job satisfaction levels occupy either high or very high. Both of these levels are approximately equal across predictors. Low and Medium environment and job satisfaction levels equally occupy the other forty percent of observations yet appear to have greater numbers of employees experiencing attrition

3.2.2.2 Continuous Variables Our continuous variables include `Income`, `Age`, `DistanceFromHome`, `TrainingTimesLastYear`, `YearsAtCompany` and `YearsWithCurrManager`. Summary statistics can be found within the appendix$_{[Fig\#6]}$. On the whole, none of these variables distributions are equally distributed. Below we mention noteworthy findings:

- `Income`: There is a narrow clustering of bars, where we find the mode of the dataset. A majority of observations fall to the left of the chart with the tail to the right, indicating it is a right-skew distribution
- `DistanceFromHome`: The distribution shows most of the employees have a short commute to work
- `TrainingTimesLastYear`: A vast majority of employees had either two or three training times
- `YearsAtCompany`: A majority of employees only have a few years of being employed at Canterra
- `YearsWithCurrManager`: The histogram shows two peaks in the dataset. The majority of employees have only been with Canterra for a few years.
- `Age`: Employee age appears slightly right skewed, indicating the average age is only slightly larger than the median age.

3.3 Correlation

Our target variable contains notably weak correlations with each of our continuous predictors$_{[Fig\#7]}$. Continuous variable distributions can be observed along the diagonal of the correlation matrix$_{[Fig\#7]}$. Though negligible in terms of magnitude, the highest correlations between our target and features were `TotalWorkingYears`, `YearsAtCompany`, `YearsWithCurrManager`, and `Age`. These features exhibited multicollinearity between one another. Normally, detecting multicollinearity may warrant the removal of one or more of these variables from the modeling process. However, due to the fact that none of these predictors exhibited sufficient magnitude with `Attrition`, we decided to include all variables for modeling. This would allow for the subsequent variance inflation values to guide an informed decision as to whether or not to remove predictors.

3.4 Feature Reduction

The `StandardHours` variable consists of no unique values other than 8. Because of the uniform occurrence of this singular value, no useful information is provided by this variable and was consequently removed from further evaluation in the modeling process.

3.5 Feature Engineering

Dummy Variable Creation - Categorical variables such as `BusinessTravel`, `Education`, `Gender`, `JobLevel`, `MaritalStatus`, `EnvironmentSatisfaction`, `JobSatisfaction`, are suited for dummy variable creation for use in our modeling. This encoding ensures these variables are appropriately interpreted by our model.

4 Model Building

4.1 Data Partition

With our goals centered on finding the most influential features in determining employee attrition, we partitioned our data 80% towards training the model and 20% for testing model performance. Allowing a significant majority of our overall employee dataset to go towards the model fitting process enables us to estimate precise, stable coefficient estimates and better overall model performance.

4.2 Logistic Regression

Our response is qualitative/categorical and therefore our problem is fundamentally one of classification and is suited for a classification model such as Logistic Regression. Logistic regression is an appropriate modeling technique for predicting employee `Attrition` due to it being a binary, categorical variable and our need to restrict the range of predicted probability values, $\mathbb{P}(Attrition)$, between 0 and 1. Our resulting probabilities can then be used to classify our observations as either the event(`Attrition` = 1) or non-event(`Attrition` = 0) given a threshold.

4.2.1 Full Model

Our initial, comprehensive logistic regression model encompasses all possible features within the Canterra employee dataset. This approach of including all features initially allows us to gain a generalized sense as to which variables exhibit the most importance in determining employee attrition.

Full Logistic Regression Model

$$\ln\left(\frac{\mathbb{P}(Attrition)}{1-\mathbb{P}(Attrition)}\right) = \hat{\beta}_0 + \hat{\beta}_1\texttt{Age} + \hat{\beta}_2\texttt{DistanceFromHome} + \hat{\beta}_3\texttt{Income} + \ldots + \hat{\beta_{27}}\texttt{JobSatisfaction\_VeryHigh}$$

Where $\ln\left(\frac{\mathbb{P}(Attrition)}{1-\mathbb{P}(Attrition)}\right)$ is the log odds of `Attrition`, $\hat{\beta}_0$ is the estimated intercept, and $\hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta_{27}}$ are the estimated coefficients of the predictor variables.

Though it may contain some useful insights, the logistic full model remains unrefined$_{[\text{Fig\#8}]}$. Nearly half of the predictors were found to be insignificant in determining the response. We did see that upon inspection of possible multicollinearity of our fitted model, no variable contributed variance inflation factors higher than a value of five, leading us to conclude that multicollinearity would not be an issue$_{[\text{Fig\#9}]}$. To refine our full logistic model, we now proceed with stepwise selection in determining our optimal set of significant predictors.

4.2.2 Reduced Model/Subset Selection

Through stepwise selection, we identified a concise number of significant predictors in determining `Attrition`$_{[\text{Fig\#10}]}$. The reduced stepwise model formula for the logit of `Attrition` can be found within the appendix$_{[\text{Fig\#11}]}$. In seeking to find a model which balances predictive power, goodness-of-fit, and complexity, our team diligently refined our stepwise selection process to base its optimal subset of predictors to minimize

the Bayesian Information Criterion(BIC)[Fig#12]. BIC, as opposed to other metrics, seeks to identify more parsimonious models, in an attempt to prevent over-complexity of models. We again found that no variance inflation factors exceeded five, leading us to conclude that multicollinearity will not effect our predictors' estimates and allowing us to be confident in their resulting significance.

4.2.2.1 Interpretation of Coefficients The coefficients for each variable individually represents the change in the log odds of `Attrition` for a one-unit change in each predictor, holding the other variables constant. For example, a one-unit increase in the `TotalWorkingYears` predictor would decrease the log odds of `Attrition` by 0.076297. Similarly, this can be interpreted as when `TotalWorkingYears` increases, the probability of `Attrition` decreases. Negative coefficients have the effect of decreasing the probability of `Attrition` whereas positive coefficients increase the probability of `Attrition`[Fig#13].

4.2.3 Performance Evaluation

To obtain a generalized sense as to the performance of both logistic models, we tested our models' performance by making predictions over the test set. For our full and final(stepwise) models, we found overall accuracy values of 85.33% and 85.57%, respectively[Fig#14/#15]. Though these appear as well-performing models, our accuracy is inflated due to imbalance in the distribution of our target variable. Many more instances of employees who did not experience attrition are prevalent. Our models have great success in correctly classifying these individuals as denoted by our specificity scores which are nearly perfect. However, our sensitivity values which estimate our models' capability in correctly identifying those individuals which will experience attrition out of all employees which do is drastically poor.

Additionally, our full and final models' Area Under the Curve (AUC) scores were 0.7513 and 0.7494, respectively[Fig#16]. AUC scores which are close to 1.0 are indicative of models which have strong discriminatory power between our class labels. The scores between our two models are nearly identically, leading us to conclude that each model is equally capable in distinguishing between employees which will and will not experience `Attrition`.

4.2.4 Recommendations to Curb Attrition

- Our stepwise logistic model coefficient estimates give us some insight as to the factors which are most important for management to address right away in order to curb attrition. Influences from `BusinessTravel`, `JobSatisfaction`, `EnvironmentSatisfaction`, and employees which are of the single `MaritalStatus` had the most magnitudal impact in determining attrition. Single employees and those which travel frequently or rarely have a greater chance of attrition than those which are not. Employees which reported to have higher levels of job and environment satisfaction had greater chances of staying with Canterra.
- Management should understand how `BusinessTravel` at both ends of the spectrum can influence the mental and physical well being of an employee. For example, too much travel may lead to feeling over-stressed, anxiety, depression, have poor sleep hygiene, a poor work-life balance, among other issues. On the other hand, employees that do not travel or travel rarely could influence other variables like an unfavorable response to `JobSatisfaction` or `EnvironmentSatisfaction`. Considering flexible business travel, like giving the employee a rest day after reaching their destination, flying during convenient hours, and/or covering additional expenses will improve the employee satisfaction and provide a greater sense of conform physically and mentally.
- The influence of `MaritalStatus` in our analysis show that single employees have the least work engagement. We can deduce single employees might have different needs than married employees. It is important for management to conduct regular surveys and feedback sessions to better understand employee needs and ambitions. For example, address opportunities for career growth, tailor mentorship programs or emphasize inclusion. Additionally, management should also monitor closely overtime work by married employees, dedicate additional resources to enhance their skills, and promote work-life balance in order to reduce the risk of attrition in this group.

4.3 Decision Trees

Because of their ease of explainability, similarity to human decision-making, and ability to capture complex, non-linear relationships, our team decided to implement classification trees. Our decision(classification) tree performs a greedy, recursive splitting where the best *split* is made based on the Gini Index. The Gini Index is a measure of the variance of classes within a node and in our case, the distribution of classes within a node. We seek to maximize the disparity of distribution between our classes(`Attrition` = 0; `Attrition` = 1) for a given split. Therefore we evaluate different splits across our feature space, resulting in a final split that minimizes the average Gini index. We trained two decision trees with the first having no controls and the second have maximum depth a complexity parameter which acts as a stopping rule.

The complexity parameter (cp) is a tuning parameter that controls the trade-off between the complexity (size) of the decision tree and its ability to fit the training data. It's a regularization parameter that helps prevent overfitting. The cp parameter demonstrates the cost of adding another variable to the tree. Higher values of cp lead to smaller trees. We have chosen our **cp value** because we believe it provides the best balance between complexity and accuracy on unseen data.

The maxdepth parameter in decision trees specifies the maximum number of levels a tree can have. As a result, the number of decisions or splits the tree can make becomes limited. Specifying the maxdepth parameter is essential for controlling the complexity of the tree and preventing overfitting. We have chosen our **maxdepth** value because we believe it provides the best balance between good performance and complexity.

Looking at our classification trees$_{[\text{Fig}\#17/\#18]}$, we see that our leaf nodes located at the bottom display the varying ways in which an employee may be classified as either one which will experience attrition or not. From the top of each tree, we see splits are being determined based on the variable and value which maximizes node/leaf purity. Following this logic, those variables which determine our splits are deemed as most significant. Based on our decision tree models, we see variables such as `TotalWorkingYears`, `Age`, `MaritalStatus_Single`, `NumCompaniesWorked`, and `Income` being among the most influential reoccurring factors affecting attrition that management can address immediately to curb attrition.

4.3.1 Peformance Evaluation

Our accuracy for both the first decision tree and second decision tree were nearly equivalent at 85.8% and 85.2%, respectively$_{[\text{Fig}\#19/\#20]}$. For the same reasons as the Logistic regression models, the accuracy metrics of these models are again inflated due to our class imbalance. As expected, the specificity value is near perfect, indicating that our model is excellent at correctly identifying employees which will not experience attrition out of those that do. However, our sensitivity values are similarly poor. Our decision tree models are not as effective when predicting employees that experience attrition out of those that do leave Canterra. Both trees are making many False Negative classifications.

The AUC for both the first and second decision tree are equally poor at approximately 0.592 and 0.583, respectively$_{[\text{Fig}\#21]}$. What this means, is that both decision trees are poor at distinguishing between our employees which will and will not leave Canterra. These decision tree AUC scores are indicative of models which are barely any better than a model which makes random predictions.

4.3.2 Recommendations to Curb Attrition

- Offering a competitive salary is the first line of defense for attrition. Compensation needs to be a priority, followed by creating a culture that employees want to be a part of. The analysis of variable `Age` shows younger employees have a higher turnover. Keeping this group in mind, we highly recommend offering attractive base salaries or hourly wages.
- Focusing on the variable `TotalWorkingYears`, we observed most employees leave the company before 2 years employment, then followed by employees leaving before 5 years of employment. In general, employees leave the company because they do not see opportunities for promotion, feel unhappy with their management, feel distanced from the organization's values or are offered a job somewhere else with higher pay.
- Observing `MaritalStatus_Single`, single employees tend to have higher attrition, by strategizing

with competitive compensation, career growth and continuous communication alongside feedback on performance, this group's retention will likely increase and have a domino effect on `TotalWorkingYears` and `Income`.

- Taking into account `NumCompaniesWorked` during the early hiring process, management could gather some insight into the employees' patterns of previous jobs and obtain feedback on what the previous companies could have done better that might apply to Canterra. For example, during the interview process, the candidate lists several jobs within a short time frame. After a conversation, the employee might specify the need of a salary increase and a more flexible schedule.

4.4 Random Forests

Recognizing that Decision Trees are subject to overfitting and high variance induced error, our team decided to implement the ensemble method of Random Forests in order to provide significant improvement over the standalone Decision Tree. Random Forests, as opposed to Decision Trees, introduce bagging (bootstrap aggregation) and feature subsetting for every split made during the model fitting process. Bagging provides diverse training sets for each tree within our Random Forest. Feature subsetting introduces a random assortment of variables of size $= \sqrt{p}$ where p is the total number of predictors that can be considered for each split in our trees. This strategy helps to not only decorrelate our trees but reduce overfitting, improve accuracy, and provide for better generalization over unseen data.

Across both of our random forest models, the variables which were found to be most important, were determined by their overall mean reduction in the Gini Index. The greater reduction in Gini Index contributed by a variable, the more important it is in the context of determining the target. The top three most influential variables found in determining attrition were `Income`, followed by `Age`, then `TotalWorkingYears`[Fig#22/#23].

4.4.1 Performance Evaluation

Our team decided to fit two robust Random Forest models. The first model is limited to one-hundred and fifty trees, while the second enforces one-thousand trees and a maximum depth of two. Random Forests with deeper depths need a higher number of trees to achieve a similar accuracy compared with shallower random forests. Both models achieved exactly the same accuracy (0.9931) and AUC value of 0.9871[Fig#24/#25/#26]. In all aspects of model performance, both random forest models performed exactly the same. In stark contrast to the other logistic and decision tree models, these random forest models have near perfect sensitivity and specificity values. A mere three employees which experienced attrition were misclassified out of all employees which left Canterra. Our team came to the conclusion that the variables which the random forest deemed to be most important, were the most trustworthy and reliable, given the overall performance of the model.

5 Overall Model Evaluation

5.1 Receiver Operating Characteristic Curve (ROC)

Receiver Operating Characteristic curves display model performance considering trade-offs between Sensitivity and 1-Specificity(False Positive Rate) at varying threshold levels and thus different confusion matrices. This represents the effectiveness in distinguishing between employees with and without attrition. The ideal ROC curve approaches the upper-left hand of the ROC plot, when we have a value of 1 for sensitivity and 0 for 1-Specificity.

The models with the most ideal ROC curves are those of the Random Forest Models[Fig#27]. Both are completely identical and occupy the upper left of the plot where the optimal ROC curve would be. These models exhibit near perfect discriminatory power in determining which employees do and do not experience attrition. The ROC curves of our two Decision Trees are nearly identical and exhibit poor discriminatory between our employees which do and do not experience attrition. This poor performance is more than likely due to our decision trees being susceptible to overfitting and variance induced error. Additionally, our full Logistic and reduced stepwise Logistic models perform similarly and exhibit a smooth curve at varying threshold values. Our Logistic models, despite being robust, are more than likely unable to capture the complex non-linear relationships inherent in our predictors relationship to employee attrition.

5.2 Area Under the Curve (AUC)

A high AUC is indicative of a model which has strong discriminatory power between our class labels. A model with perfect discriminatory power has an AUC value of 1. A model which makes random predictions is essentially worthless and has an AUC value of 0.5. The models that undoubtedly are the most effective in distinguishing between employees with and without attrition are the Random Forest Models. This can be determined by comparing each of our models' Area Under the Curve (AUC) values. Our Random Forest models are the highest at a near perfect value of 0.9871. Our next highest AUC values come from our logistic models. Despite our full logistic model and stepwise logistic model not being as optimal as our random forests', they do exhibit a moderate level of discriminatory power at AUC values of .7513(full) and .7494(stepwise). Lastly, our decision tree models exhibit poor performance at 0.5925(no parameters) and 0.5832(maxdepth/cp parameters).

6 Conclusion

6.1 Optimal Model Decision

Our team deliberated on multiple types of models, including logistic regression, decision trees, and random forests in an attempt to decipher the key employee demographics which were most significant in determining employee attrition. While our logistic regression and decision tree models offered explainability, they failed to achieve an adequate level of predictive performance in capturing the underlying, complex relationships inherent in our employee data. Such lack of performance would not inspire confidence in Canterra's predictions as to which employees are most likely to experience attrition.

However, the random forest models attained such a high level of predictive performance, that our team felt confident in presenting either to Canterra. Despite both random forest models achieving the exact same performance, we have opted to use the random forest with one-thousand trees and a max depth of two in presenting our results to Canterra. The shallow depth of the random forest trees, help prevent overfitting, but demand a greater number of trees to achieve optimal performance. We are choosing to put forward this model to Canterra, as we believe that it is more robust given that there are more trees included in the overall performance of the model. Our proposed random forest model achieved an accuracy of 99.31% and AUC 0.9871. In stark contrast to the other logistic and decision tree models, our proposed random forest model has near perfect sensitivity and specificity values. The sensitivity, which is in particular interest to Canterra, denotes the ability of our model to correctly predict the employees that leave Canterra out of those that actually do. With such drastically high performance, our proposed random forest model provides Canterra with additional confidence in the aforementioned important factors in determining attrition.

6.2 Recommendations

Given our findings on some of the most influential predictors such as `BusinessTravel`, `JobSatisfaction`, and `EnvironmentSatisfaction`, Canterra would find it most productive to focus on improving in these areas. Reducing business travel to only necessary classifications would increase job satisfaction and overall reduce attrition among those who currently travel frequently. From the perspective of our random forest model, the most important factor in determining employee attrition was `Income`. We recommended that Canterra reevaluate its salary ranges for young employees in particular. An increasingly common practice in the current market is that young employees find job-hopping between different companies provides significant bumps in their salary. If Canterra can attempt to salary match between its current competitors, employees will be more willing to remain at Canterra. `Age` was found to be the second most important factor in determining attrition. It would prove fruitful to increase efforts around young, single employees who have not been with the company for very long. It would be wise to implement programs and initiatives to help these employees feel a sense of belonging and encourage a high job satisfaction level. Our analysis has shown that the older the employee is and the longer they have been with the company, they are less likely to leave. Keeping young employees' job satisfaction levels high as they age, encourages them to stay with the company longer, and as a result are even less likely to leave.