**Case Study 2: People Analytics**

**Group 11:**

**Benjamin Fender**

**Lee Green**

**James Grissel**

**March 30, 2020**

**BAN840 – SP20**

**Sharma**

**Executive Summary**

**People Analytics – Using Information to Reduce Employee Churn**

A key question in business today is how to achieve the highest return for the company’s investment dollars. Analysis has continually focused on the production of vital products and services, the retention and acquisition of customers, and the efficiency of internal processes and functions. While there are some examples of workforce management since the industrial revolution, most human resource departments have continually been considered operational and old-fashioned (Vulpen, 2017). In most companies, the HR department is in a transition. The increase of computer power and the onset of the internet led to the birth of the modern age of HR Analytics in 1991 (MicroStrategy, 2020). Companies harnessed the power of their HR management information systems and curated the analytical lessons learned by their sales teams and shop floor intelligence efforts to bolster their most valuable asset, their people.

Strategic Analytical Human Resource planning has become a broad field that successful companies look to for insights to the get the most out of their employees. An industry accepted definition for the practice is: “HR Analytics is the systematic identification and quantification of the people drivers of business outcomes” (Heuvel & Bondarouk, 2016). Why does an employee leave their company? What level of happiness convinces them stay? What is a good metric for employee satisfaction and happiness? Is it an emotional decision, or are there key contributors to the decision that can help in predicting the turnover before it happens?

While employee satisfaction is often recognized as a value-add to the company, many groups are starting to measure the impact of employee loss through analytics. In many fields it is hard to have a uniform measure for employee loss, but some studies have made an effort to calculate it. The average cost to hire an employee across all industries was $4,129 in 2016 (SHRM, 2016) and the cost of employee loss for an average employee (those with median salaries of $45K/Year) is $15,000 (HRDive, 2017). Companies would rather retain good employees than incur the cost of hiring replacements.

The stakes only rise when the attrition is incurred by a manager. Manager loss not only carries a cost impact, but there is a new measure of revenue-at-risk due to the change. The industry-average productivity loss is $845K and $3.5M in annual revenue at risk. These calculations use the following assumptions: a company average of 10,000 employees, an average span of control of 7 employees per manager, a 5.4% resignation rate of managers per year, industry benchmark average manager salary and revenue-per-employee to the Visier Manager Turnover Calculator (Visier, 2020).

In the desire for companies to retain employees, companies are looking for new ideas. Mary Hogen collected the following measures in her article on employee retention keys (Hogen, 2015):

* *One third of new hires quit their job after about six months*
* *Referred employees have a 45 percent retention rate after two years*
* *Remote workers are 50 percent less likely to quit*
* *One third of employees knew whether they would stay with their company long-term after their first week*
* *Some 35 percent of employees will start looking for a job if they don’t receive a pay raise in the next 12 months*

The metrics listed above are examples a mix of traditional analytics and newer novel measures. These metrics outline the importance of retaining employees to save a company thousands in lost profits. Our analysis will walk you through various factors that contribute to attrition. While not all of the information used by other studies is available to us, we will use dependable techniques and key information to create actionable insights for the management of the workforce.

**Predicting Employee Attrition**

Identifying the factors why employees voluntarily or involuntarily left a company are important to a human resources (HR) department of an organization. However, determining the factors of employee attrition provide more value for an HR department for the retention of current employees. This shift in analysis delivers a prediction of what types of employees are most likely to leave a company before the action. Also, the company can make the necessary adjustments to preserve the investments in money and time spent on each employee to increase the probability that the employee will stay. The following analysis utilizes logistic regression techniques to gain insight into the factors for attrition and how to apply those factors to retain current employees.

**HR Dataset**

The dataset used in the analysis includes attributes for 1,470 employees from a company. The attributes include personal and professional information normally monitored by an HR department, including an Attrition variable that identifies if an employee left his or her position or not. The dataset will be evaluated to determine what degree all or part of this data predicts an employee’s future retention status. After completing dataset exploration, the following was determined:

* The dataset appears to be complete with no missing values, so we did not need to impute any information
* The data is structured to include one-row per employee. Within the rows there is textual data and whole integers.
* 25% of the fields are discrete values (categorical) and 75% of the fields are continuous data.
* Some fields have equivalent data for all observations, so those are likely candidates to be removed from the analysis.
* At least one, field is likely insignificant for predictive analysis (EmployeeNumber) as it appears to be a semi-sequential number used to identify individual employees
* There does appear to be some potential ordinal data in the dataset. BusinessTravel and MaritalStatus strongly present as ordinal data.

*(Include data structure information, categorical variables, nominal vs ordinal, etc.)*

**Descriptive Statistics**

The descriptive statistics for the Employee Attrition dataset was derived with Regressit via Excel. These statistics provided summaries of samples and measures. In reviewing the output we will be looking at measures of central tendency and measures of variability (spread). (Kenton, 2019).

Measures of Variability: We identified (Table 1) seven fields (29%) that showed to be extremely right skewed with a positive value above 1.0, an additional seven fields (29%) displayed a skewness value between .5 and 1 and where moderately skewed. The final 10 fields (42%) had a much more normal skewness value of less than .5. Two fields that had all duplicated data did not have usable skewness statistics. The fields with a high skew value will likely need to be transformed prior to model trials. Two fields (YearsAtCompany and YearsSinceLastPromotion), shown in Graph 1A&B, displayed long tails and had a Kurtosis Value above the normal distribution 3.0

Measures of Central Tendency: 10 (35%) of the fields have a standard deviation that is 50% or more higher than the median. This likely indicates less central tendency and more spread in these fields. We will need to compare the standard error and the histograms of the fields to determine the data structure. Additionally, there is a pattern in the minimum values. More than 50% of the data set has minimum values of 0 or 1. This helps confirm our data exploration observations of potential ordinal values and the rounding of values, especially those less than 1.

Review of Multicollinearity: The primary focus of reviewing the correlation matrix (Table 2) was the Age Field. While no field has a correlation approaching 1, Age has a higher correlation to JobLevel, MonthlyIncome, NumCompaniesWorked, and not surprisingly TotalWorkingYears, These values appear to go up with age.

**Data Cleaning and Preparation**

In preparation and cleansing of the dataset for analysis, exploration observations related to the types of variables, attributes to be removed, and the overall structure of the dataset were acted on.

* The following attributes contained the same data for each record and were removed from the dataset: Over18, Standard Hours, and EmployeeCount.
* BusinessTravel appeared to be an ordinal value. It was recoded as BusinessTravel.Ord with Non-Travel=0, Travel\_Rarely=1, and Travel\_Frequently=2.
* MaritalStatus, was explored to see if an ordinal relationship existed and added power to the model. However, it was found through trials that MaritalStatus = Single was the strongest predictor.
* Several of the fields appeared to have a left-skew and in the process of the trials it was deemed necessary to transform the data with a natural log. Additionally, much of the numerical data that is focused on time is in integer form and <1 is 0. To accommodate the natural log transformation, we added a consistent value (2) to all of the observations where 0 was contained in the observations.

**Results**

Following multiple trials and testing, a logistic regression was calculated the predicted probability of Employee Attrition being positive based on the 10 features selected in the model (Table 3). The equation below has an **R-Squared of 0.261 and Adj. R-Sqr of 0.244**

Predicted probability of "Attrition.Eq.Yes = Yes" is equal to exp(LogOdds)/(1+exp(LogOdds)) = 1/(1+exp(-LogOdds))

where LogOdds = 5.437 + 0.854\*BusinessTravel.Ord + 0.041\*DistanceFromHome - 0.398\*EnvironmentSatisfaction - 0.59\*JobInvolvement - 0.371\*JobSatisfaction + 0.971\*MaritalStatus.Eq.Single - 0.468\*MonthlyIncome.Ln + 0.171\*NumCompaniesWorked + 1.703\*OverTime.Eq.Yes - 1.075\*TWYAdj.Ln, where Attrition is coded as 1 = Yes, 0 = No

Variable Selection: To perform variable selection, we utilized the caret model in R to run GLM logistic regression analysis. Along with the regression, we used the ANOVA and Variable Importance features in caret to identify influential variables. From there, we used RegressIt in Excel to perform additional variable transformation and step-wise regression methods to tune the model using adjusted R-squared, RMSE, and parsimony to guide our decision-making.

Cross-Validation: The 5-fold cross-validation technique was used. Raheel Sahikh defines this type of validation as so: The splitting of data into folds may be governed by criteria such as ensuring that each fold has the same proportion of observations with a given categorical value, such as the class outcome value. This is called stratified cross-validation. (Shaikh, 2018). To perform cross-validation, we first generated stratified folds in R using the folds function in the mltools module. We then wrote those folds back to Excel to use in RegressIt.(Table 6) From there, we created sample columns in the Excel data using the in-sample data for each fold. Then, we ran regressions on each fold using the previously established model, using the entire response variable data as our out-of-sample data.

Selection of appropriate performance criteria and cutoff level: For our performance criteria, we believe that Accuracy is a reliable all-around measure and that Sensitivity is a logical measure given our perceived importance of True Positive Rate. In an ideal situation, we would have an idea of the costs of False Positives and False Negatives.

To choose our cutoff level we first made visual inspections of the ROC curves (Graph 2) as well as the Distribution of Outcomes -vs- Prediction Interval Chart (Graph 4). That provided an initial indication that we should explore the 0.50 cutoff level. Next, we used our cross-validation sets (Table 6) to calculate the performance of each fold at several cutoff levels (0.35-0.6) at 0.05 intervals(Table 7) We ultimately selected 0.55 as a good balance between Accuracy and Sensitivity highlighted in the Sensitivity and Specificity -vs- Cutoff Value (Graph 3)

**Summary**

The predicted probability of a true positive for Attrition.Eq.Yes = Yes presents 72 of the 1470 employees leaving the company, with a cutoff value of 0.55. The predicted probability of a true negative represents 1213 of 1470 employees. The true outcomes that were correct is 87.4% of the dataset. Table 4 shows the confusion matrix and Table 5 shows a sample of the model employee predicted attrition based on the model logistic regression equation.

An HR department can use this model to predict employee attrition and how to keep the company’s investments intact. Adjusting the variables that influence the equation for employees that are predicted to leave the company can show what needs to be done for valuable employees to stay. For example, a decrease in job satisfaction increases the probability for attrition. Also, when hiring a new employee, the greater number of companies worked for increases the probability for attrition in the future.

**Appendix**

Table 1 – Measure of Variability – Skewness and Kurtosis



Graph 1 A&B – Kurtosis Above 3

Table 2 – Correlation Matrix



Table 3 – Logistic Regression Coefficient Estimates

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std.Err.** | **z-statistic** | **P-value** | **Lower95%** | **Upper95%** | **VIF** | **Std. coeff.** |
| Constant | 5.437 | 1.470 | 3.698 | 0.000 | 2.555 | 8.318 |  |  |
| BusinessTravel.Ord | 0.854 | 0.159 | **5.364** | 0.000 | 0.542 | 1.166 | 1.006 | 0.250 |
| DistanceFromHome | 0.041 | 0.009969 | **4.157** | 0.000 | 0.022 | 0.061 | 1.004 | 0.185 |
| EnvironmentSatisfaction | -0.398 | 0.076 | **-5.243** | 0.000 | -0.547 | -0.249 | 1.006 | -0.240 |
| JobInvolvement | -0.590 | 0.115 | **-5.119** | 0.000 | -0.815 | -0.364 | 1.005 | -0.231 |
| JobSatisfaction | -0.371 | 0.075 | **-4.937** | 0.000 | -0.518 | -0.224 | 1.005 | -0.226 |
| MaritalStatus.Eq.Single | 0.971 | 0.169 | **5.748** | 0.000 | 0.640 | 1.302 | 1.018 | 0.250 |
| MonthlyIncome.Ln | -0.468 | 0.197 | -2.370 | 0.018 | -0.854 | -0.081 | 2.236 | -0.171 |
| NumCompaniesWorked | 0.171 | 0.034 | **5.076** | 0.000 | 0.105 | 0.237 | 1.092 | 0.236 |
| OverTime.Eq.Yes | 1.703 | 0.174 | 9.804 | 0.000 | 1.362 | 2.043 | 1.009 | 0.423 |
| TWYAdj.Ln | -1.075 | 0.204 | **-5.278** | 0.000 | -1.475 | -0.676 | 2.361 | -0.361 |

Table 4 – Confusion Matrix



Table 5 – Predicted Probabilities



Graph 2

Graph 3

Graph 4

Table 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cutoff Level | Accuracy | Sensitivity (TPR) | Specificity (TNR) |
| 1 | 0.45 | 86.7% | 25.5% | 98.4% |
| 1 | 0.4 | 85.4% | 31.9% | 95.5% |
| 1 | 0.35 | 85.4% | 40.4% | 93.9% |
| 2 | 0.45 | 89.1% | 56.3% | 95.5% |
| 2 | 0.4 | 89.5% | 64.6% | 94.3% |
| 2 | 0.35 | 88.8% | 68.8% | 92.7% |
| 3 | 0.45 | 89.1% | 53.2% | 96.0% |
| 3 | 0.4 | 89.1% | 57.4% | 95.1% |
| 3 | 0.35 | 88.8% | 66.0% | 93.1% |
| 4 | 0.45 | 90.1% | 54.2% | 97.2% |
| 4 | 0.4 | 90.1% | 58.3% | 96.3% |
| 4 | 0.35 | 88.8% | 60.4% | 94.3% |
| 5 | 0.45 | 88.1% | 36.2% | 98.0% |
| 5 | 0.4 | 89.1% | 46.8% | 97.2% |
| 5 | 0.35 | 86.4% | 46.8% | 93.9% |
| 1 | 0.55 | 86.4% | 21.3% | 98.8% |
| 2 | 0.55 | 89.1% | 45.8% | 97.6% |
| 3 | 0.55 | 89.8% | 38.3% | 99.6% |
| 4 | 0.55 | 88.8% | 43.8% | 97.6% |
| 5 | 0.55 | 88.1% | 31.9% | 98.8% |
| 1 | 0.5 | 86.1% | 21.3% | 98.4% |
| 2 | 0.5 | 90.1% | 52.1% | 97.6% |
| 3 | 0.5 | 88.1% | 40.4% | 97.2% |
| 4 | 0.5 | 89.1% | 47.9% | 97.2% |
| 5 | 0.5 | 88.1% | 34.0% | 98.4% |
| 1 | 0.6 | 86.1% | 14.9% | 99.6% |
| 2 | 0.6 | 89.5% | 45.8% | 98.0% |
| 3 | 0.6 | 89.8% | 36.2% | 100.0% |
| 4 | 0.6 | 88.1% | 37.5% | 98.0% |
| 5 | 0.6 | 88.1% | 29.8% | 99.2% |

Table 7

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Column Labels |  |  |  |  |  |
| Values | **0.35** | **0.4** | **0.45** | **0.5** | **0.55** | **0.6** |
| Average of Accuracy | 87.6% | 88.6% | 88.6% | 88.3% | 88.4% | 88.3% |
| Average of Sensitivity (TPR) | 56.5% | 51.8% | 45.1% | 39.1% | 36.2% | 32.8% |
| Average of Specificity (TNR) | 93.6% | 95.7% | 97.0% | 97.7% | 98.5% | 98.9% |

# References

Heuvel, S. v., & Bondarouk, T. (2016). The Rise (and Fall) of HR Analytics. *Article submitted for the 2nd HR Division International Conference (HRIC).* Sidney, Australia.

Hogen, M. (2015, November 30). *9 Employee Retention Statistics That Will Make You Sit Up and Pay Attention*. Retrieved from TLNT Talent Management & HR: https://www.tlnt.com/9-employee-retention-statistics-that-will-make-you-sit-up-and-pay-attention/

HRDive. (2017). *Study: Turnover costs employers $15,000 per worker*. Retrieved from hrdive.com: https://www.hrdive.com/news/study-turnover-costs-employers-15000-per-worker/449142/

Kenton, W. (2019, July 19). *Corporate Finance & Accoutning*. Retrieved from Investopedia: https://www.investopedia.com/terms/d/descriptive\_statistics.asp

MicroStrategy. (2020). *HR Analytics - Everything You Need to Know*. Retrieved from MicroStrategy.com: https://www.microstrategy.com/us/resources/introductory-guides/hr-analytics-everything-you-need-to-know

Plummer, K. (2014, October 2). *Linkedin SlideShare*. Retrieved from LinkedIn: https://www.slideshare.net/plummer48/reporting-a-multiple-linear-regression-in-apa

Shaikh, R. (2018, November 26). *towards data science*. Retrieved from torwardsdatascience.com: https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

SHRM. (2016, August 3). *Average Cost-per-Hire for Companies Is $4,129, SHRM Survey Finds*. Retrieved from SHRM.org: https://www.shrm.org/about-shrm/press-room/press-releases/pages/human-capital-benchmarking-report.aspx

Visier. (2020, March). *What does Manager Turnover cost you?* Retrieved from visier.com: https://www.visier.com/manager-turnover-calculator/

Vulpen, E. v. (2017). *What is HR Analytics?* Retrieved from AIHR Analytics: https://www.analyticsinhr.com/blog/what-is-hr-analytics/