CAUSAL INFERENCE OF HUMAN RESOURCES KEY PERFORMANCE INDICATORS

Matthew Kovach

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTERS OF SCIENCE

December 2018

Committee:

Wei Ning, Advisor

John Chen

Junfeng Shang

ABSTRACT

Wei Ning, Advisor

The purpose of this study is to examine the relationship between attrition rates and key performance indicators in a corporate workforce by using the propensity score (PS) matching. The study shows the possibilities of using logistic regression and propensity score matching methods in human capital strategic decisions. The data used here was from a fictional data set created by IBM data scientists based on active and separated employees to uncover the factors that lead to employee attrition. For each of the 1,470 employee records, information was generated about demographic characteristics such as age, gender, marital status, education level, employment status and culture, compensation, and performance factors. [1]

Two logistic equations are defined for two key performance objectives, culture and work life balance. A logistic regression analysis on each equation, with support from contrast estimation, reveals a comparison between the most and least favorable responses to key performance indicators is most significant. After successfully balancing a treatment and control group using the nearest neighbor matching technique on propensity score estimates from the logistic regression, a paired t-test reveals a statistically significant difference for the work life balance key performance indicator. This result is interpreted as having the highest probability of successfully reducing attrition when the focus is on increasing employee responses to satisfaction levels of work life balance in comparison to other key performance indicators.

[1] https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Wei Ning for his continuted support as an advisor, a teacher and a friend. His guidance throughout my academic career has certainly helped define my future personal and professional goal. I would also like to acknowledge Dr. Craig Zirbel and the Department of Mathematics and Statistics for their support in helping me achieve my educational goals at Bowling Green State University.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER 1 INTRODUCTION

Many companies have implemented tools for measuring their human capital performance in order to stay competitive with global labor markets. Organizations are forced to measure human capital performance and contribute to the stability of the organization's human capital structure. In other words, organizations are adopting a data-driven approach towards human resource management.

Human resource management has changed. It has moved from an operational discipline towards a more strategic discipline. Part of the strategic discipline includes implementing key performance indicators, also known as KPIs, to help managers and employees gauge the effectiveness of various functions and processes important to achieving organizational goals.[1]

Definition 1.0.1. ***Key Performance Indicator***

Key performance indicators (KPI) are a set of quantifiable measures that a company uses to gauge its performance over time. These metrics are used to determine a company's progress in achieving its strategic and operational goals, and also to compare a company's finances and performance against other businesses within its industry.

Definition 1.0.2. ***Key Performance Objectives***

Key performance objectives (KPO) are a set of related KPIs that describe a company's specific strategic or operational goal.

The objective is to identify those measures that meaningfully communicate accomplishment of or progress toward key performance objectives. Without adequate data over time, many companies rely on their observational data to draw causal inferences on key performance indicators. The issue becomes finding two groups of employees to make comparisons and draw inferences.

In observational studies, the groups compared are often different because of the lack of randomization. Subjects with specific characteristics may be more likely to affect the outcome variable

[1] Organizational goals in this paper are also referred to key performance objectives. Each key performance objectives has a subset of related key performance indicators.

than other subjects. If these characteristics also affect the outcome, a direct comparison of the groups is likely to produce biased conclusions that may merely reflect the lack initial comparability.[2] Logistic regression is a commonly used method to control for imbalances between groups. Its primary advantage is the ability to control for many variables simultaneously. Another method to control for imbalances is the propensity score, which is the conditional probability of a subject receiving a particular treatment given the set of confounders.

Propensity score matching (PSM) was first introduced by Rosenbaum and Rubin (1983) in "The Central Role of the Propensity Score in Observational Studies for Casual Effects." Propensity scores offer an alternative method to estimate the effect of receiving treatment when random assignment of treatments to subjects is not feasible. PSM refers to the pairing of treatment and control units with similar values on the propensity score, and possibly other covariates (the characteristics of the population), to remove the selection bias between the treatment and control groups. Like other matching procedures, propensity score matching estimates an average treatment effect from observational data. This matching can help strengthen the causal arguments in observational studies. Some of the benefits associated with propensity scores are: (a) creating adequate counterfactuals when random assignment is infeasible, or when the interest is in assessing treatment effects from survey, administrative, or other types of data where treatment assignment is uncontrollable and (b) reducing the number of covariates needed to control for unexplained variances.

The general procedure for the paper is as follows: (1) run logistic regression, (2) match observations on propensity score with nearest neighbor matching, and (3) conduct an outcome analysis based on new sample of matched propensity scores. Chapter 2 introduces the dataset in more detail along with a logistic regression analysis of the key performance objectives culture and work life balance. Two logistic regression equations are proposed for each key performance objectives along with balancing covariates to account for individual differences and reduce variability. Chapter 3 discusses the rationale for implementing propensity score matching and tests for differences in covariate means between the control and treated groups of the nearest neighbor matching technique.

[2]These characteristics are called cofounders.

An outcome analysis is performed in Chapter 4 to test the difference in attrition means using the t-test and paired t-test on significant contrasts found in Chapter 2 in order to determine which key performance indicator has the highest probability of successfully reducing attrition. Chapter 5 concludes the thesis with a discussion of the work.

CHAPTER 2 LOGISTIC ANALYSIS OF KEY PERFORMANCE OBJECTIVES

The human resources employee attrition experimental dataset consists of multiple sets of key performance objectives with many related key performance objectives. The focus of this chapter i is to examine the KPOs related to organizational culture and work-life balance. The key performance indicators related to organizational culture are job satisfaction and environment satisfaction whereas work-life balance includes the key performance indicators work life balance and distance. In the experimental data, there are four measures about the key performance indicators, whether employees have record their satisfaction levels as 1=Low, 2=Medium, 3=High, or 4=Very High.[1] The two possible dependent variable levels here represent retention; the individual is either still active or has voluntarily or involuntarily exited the company. The variable attrition is labeled as "0" for remains active and "1" for no longer with the company. Table 3.1 shows the rate of attrition for the entirety of the workforce. Tables 2.2 - 2.4 represent the rate of attrition by employee response to the key performance objectives culture and work life balance.

Table 2.1 Summary Statistics for Count of Attrition

Attrition	
No (0)	1233
Yes (1)	237

Table 2.2 Rate of Attrition for Job Satisfaction

Job Satisfaction	Employees	Attrition %
Low	289	0.228
Medium	280	0.164
High	442	0.165
Very High	459	0.113

Figure 2.1 shows the rate of attrition for each key performance indicator; job satisfaction, environment satisfaction and work life balance. The rating levels indicate whether the employee

[1]The KPI distance is measured in driving miles from home to work.

Table 2.3 Rate of Attrition for Environment Satisfaction

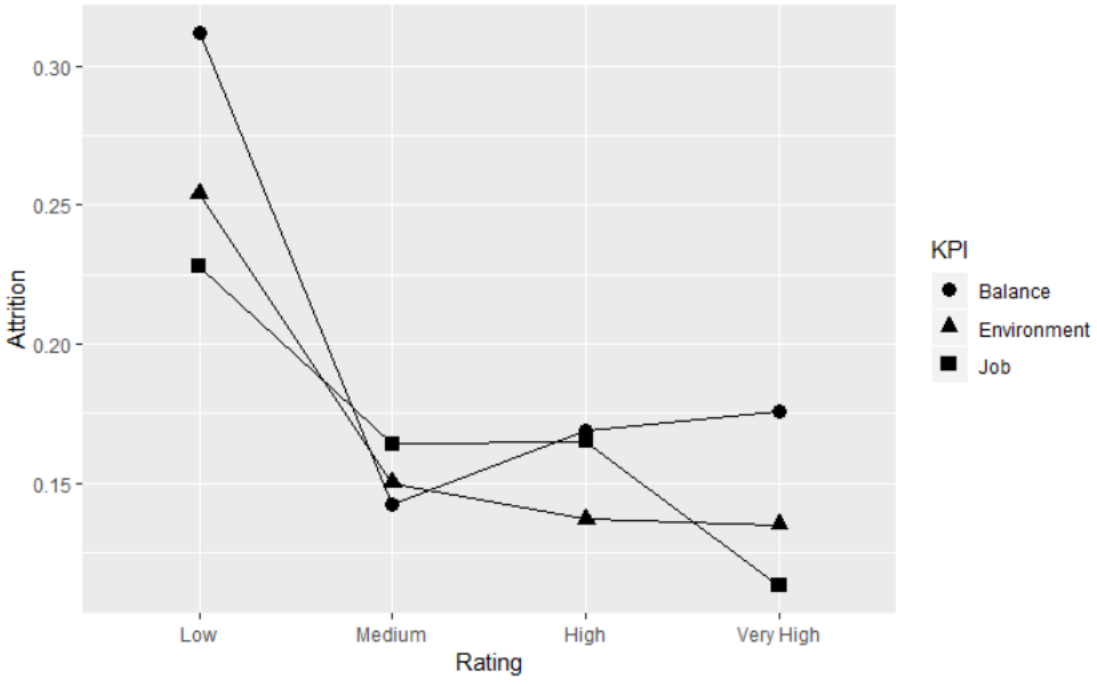
Environment Satisfaction	Employees	Attrition %
Low	284	0.254
Medium	287	0.150
High	453	0.137
Very High	446	0.135

Table 2.4 Rate of Attrition for Work Life Balance

Work Life Balance	Employees	Attrition %
Bad	80	0.312
Better	893	0.142
Good	344	0.169
Best	153	0.176

responded with (1) low satisfaction, (2) medium satisfaction, (3) high satisfaction, or (4) very high satisfaction. [2]

Figure 2.1 Rate of Attrition



[2] A rating level of 4 is considered most favorable; a rating level of 1 is considered least favorable.

2.1 Logistic Regression

The logistic model is a statistical model that is usually taken to apply to a binary dependent variable. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of predictor variables. The goal of a logistic regression is to describe the relationship between the dichotomous characteristic of attrition and a set of predictor variables containing measures of related key performance indicators and a set of benchmark employee characteristics to achieve a reduction in variability. A set of logistic equations are defined to model the two key performance objectives. Each logistic equation includes the set of KPIs along with education, department, marital status, gender, and age as benchmark variables:

Organizational Culture

$$logit(p) = \beta_0 + \beta_1 JobSatisfaction + \beta_2 EnvironmentSatisfaction + \beta_3 education + \beta_4 department + \beta_5 maritalstatus + \beta_6 gender + \beta_7 age \quad (2.1.1)$$

Work-Life Balance

$$logit(p) = \beta_0 + \beta_1 WorkLifeBalance + \beta_2 Distance + \beta_3 education + \beta_4 department + \beta_5 maritalstatus + \beta_6 gender + \beta_7 age \quad (2.1.2)$$

A logistic regression analysis is then generated for each equation independently to predict a logit transformation of the probability of attrition, p.

2.1.1 Analysis of Culture Key Performance Objective

A global likelihood ratio test to measure how well the observed data corresponds to the fitted model is given in Table 2.5. The model is a good fit to the observed data; the null model is rejected in favor of the alternative model at a nominal level $\alpha = 0.05$.

The key performance indicators in Table 2.6 for the key performance objective, culture, are

Table 2.5 Logistic Regression Model Fit

Global Test	Chi-Square	DF	Pr>ChiSq
Likelihood Ratio	129.5804	16	<.0001
Score	128.6738	16	<.0001
Wald	114.4533	16	<.0001

each statistically significant. The output indicates that job satisfaction and environment satisfaction and all covariates but education and gender are significantly associated with the probability of attrition. Forward selection methods generally agree with the inclusion of all covariates, however, statistically insignificant covariates are still included in order to compare the two sets of logistic equations.[3]

Table 2.6 Logistic Regression Analysis of Effects

Effect	DF	Chi-Square	Pr>ChiSq
Job Satisfaction	3	19.7013	0.0002
Environment Satisfaction	3	23.0491	<.0001
Education	4	0.86330	0.9298
Department	2	10.9568	0.0042
Marital Status	2	36.8248	<.0001
Gender	1	1.77360	0.1829
Age	1	26.6261	<.0001

The odds ratio in Table 2.7 gives the relative amount by which the odds of attrition increase or decrease when the value of one key performance indicator for culture is increased by one unit.[4] An odds ratio is a relative measure of effect, which allows the comparison of a treatment and control group. If the outcome is the same in both groups, the ratio will be one, which implies there is no difference between the two groups. The output below shows statistically significant estimates only. For example, the odds of attrition for those employees with low job satisfaction are 0.409 times as large in comparison to the odds of attrition for employees with medium job satisfaction.

Interpreting the odds ratio may provide misleading evidence. The frequency of attrition varies among the levels of satisfaction for each key performance indicator. It is best to find alternative

[3]See Appendix C for the AIC and maximum likelihood estimates for the logistic equations.

[4]This paper does not attempt to interpret covariate estimates; covariates are included to reduce variability and provide more accurate comparisons.

methods for the purpose of interpretation. In order to verify the significant of the odds ratio estimates, as well as determine the most significant employee response comparison to further analyze, an appropriate list of contrasts is constructed. See Table 2.8 and 2.9 below. [5]

Table 2.7 Logistic Regression Odds Ratio

Effect		Estimate	95% Wald Confidence Limits	
Job Satisfaction	Low vs Medium	0.409	0.991	0.409
	Low vs Very High	0.255	0.589	0.255
	Medium vs Very High	0.609	0.389	0.953
	High vs Low	1.065	2.336	1.065
	High vs Very High	0.612	0.410	0.911
Environment Satisfaction	Low vs Medium	0.488	0.314	0.758
	Low vs. Very High	0.286	0.638	0.286
Department	R&D vs Sales	1.636	1.200	2.229
Marital Status	Divorce vs Single	2.977	1.936	4.580
	Married vs Single	2.290	1.657	3.165
Age		1.048	1.030	1.067

Table 2.8 Model Fit Statistics for Contrast Estimation

Contrast	Chi-Square	DF	Pr>ChiSq
Job Satisfaction	19.7013	3	0.0002
Environment Satisfaction	23.0431	3	<.0001

Note the following explanations for the mean comparison of contrast estimates. Very High // Low iis the mean difference between employees with very high satisfaction and low satisfaction. High // Low is the mean difference between employees with high satisfaction and low satisfaction. Upper // Lower is the difference between means of employees with very high and high satisfaction and those with low and medium satisfaction.

The output finds a statistically significant difference in means among the four levels of responses. Additionally, the most significant contrast estimate for both key performance indicators is the mean difference between very high satisfaction and low satisfaction. [6]

[5] See Appendix C for the contrast confidence intervals.

[6] Both estimates have the largest chi-sq value along with the smallest rejection region.

Table 2.9 Contrast Estimates

Contrast		Estimate	Std. Error	Chi-Sq	Pr>ChiSq
Job Satisfaction	Very High // Low	1.5899	0.2063	12.771	0.0004
	High // Low	1.5775	0.3161	5.1773	0.0229
	Upper // Lower	1.2481	0.1558	3.1530	0.0758
Environment Satisfaction	Very High // Low	1.8195	0.2322	22.000	<.0001
	High // Low	2.2841	0.4656	16.417	<.0001
	Upper // Lower	1.6040	0.2005	14.297	0.0002

2.1.2 Analysis of Work-Life Balance Key Performance Objective

A global likelihood ratio test to measure how well the observed data corresponds to the fitted model is given in Table 2.10. The model is a good fit to the observed data; the null model is rejected in favor of the alternative model at a nominal level $\alpha = 0.05$.

Table 2.10 Logistic Regression Model Fit

Global Test	Chi-Square	DF	Pr>ChiSq
Likelihood Ratio	114.2579	14	<.0001
Score	114.5467	14	<.0001
Wald	102.5012	14	<.0001

The key performance indicators in Table 2.11 for the key performance objective, work life balance, are each statistically significant. The output indicates that work life balance and distance and all covariates but education and gender are significantly associated with the probability of attrition. Forward selection methods generally agree with the inclusion of all covariates. [7]

Table 2.11 Logistic Regression Analysis of Effects

Effect	DF	Chi-Square	Pr>ChiSq
Distance	1	10.687	0.0011
Work Life Balance	3	18.128	0.0004
Education	4	0.6455	0.9579
Department	2	11.717	0.0029
Marital Status	2	37.606	<.0001
Gender	1	1.6539	0.1984
Age	1	25.960	<.0001

[7] See Appendix C for the AIC and maximum likelihood estimates for the logistic equations.

The odds ratio in Table 2.12 gives the relative amount by which the odds of attrition increase or decrease when the value of on key performance indicator for work-life balance is increased by one unit.[8] The output below shows statistically significant estimates only. For example, the odds of attrition for those employees with bad work life balance satisfaction are 0.418 times as large in comparison to the odds of attrition for employees with good work life balance satisfaction.

In order to verify the significance of the odds ratio estimates, as well as determine the most significant employee response comparison to further analyze, an appropriate list of contrasts is again constructed.. See Table 2.13 and 2.14 below.[9] The output finds a statistically significant difference in means among the four levels of responses and suggests the mean comparison of best vs bad is most different. [10]

Table 2.12 Logistic Regression Odds Ratio

Effect		Estimate	95% Wald Confidence Limits	
Work Life Balance	Bad vs Good	0.418	0.234	0.748
	Bad vs Better	0.317	0.185	0.543
	Bad vs Best	0.417	0.215	0.809
Department	R&D vs Sales	1.662	1.221	2.261
Marital Status	Divorced vs Single	2.912	1.897	4.470
	Married vs Single	2.343	1.698	3.233
Age		1.048	1.029	1.067

Table 2.13 Model Fit Statistics for Contrast Estimation

Contrast	Chi-Square	DF	Pr>ChiSq
Work Life Balance	18.1280	3	0.0004

Note the following explanations for the mean comparison of contrast estimates. Best // Bad is the mean difference between employees with best satisfaction and bad satisfaction responses. Good // Bad is the mean difference between employees with good satisfaction and bad satisfaction

[8]This paper does not attempt to interpret covariate estimates; covariates are included to reduce variability and provide more accurate comparisons.

[9]Upper vs. Lower is defined as the average difference between the top responses and bottom responses.

[10]See Appendix C for the contrast confidence intervals.

responses. Upper // Lower is the difference between means of employees with best and good satisfaction responses and those responding with bad and better satisfaction.

Table 2.14 Contrast Estimates

Contrast		Estimate	Std. Error	Chi-Sq	Pr>ChiSq
Work Life Balance	Best // Bad	2.3998	0.8122	6.6896	0.0097
	Best // Better	0.7596	0.1842	1.2851	0.2569
	Upper // Lower	1.2521	0.2170	1.6822	0.1946

The output finds a statistically significant difference in means among the four levels of responses. Additionally, the most significant contrast estimate for the work life balance key performance indicators is the mean difference between best satisfaction and bad satisfaction. [11] Here we were able to adequately describe the rate of attrition based on specific key performance objectives and a set of benchmark characteristics using a logistic regression. This allows us to interpret the mean differences in attrition by employee responses to the key performance indicators. Although each difference in means can be further analyzed, the thesis focuses next on the most significant contrast estimate to provide further analyses and recommendations.

[11] The estimate has the largest chi-sq value along with the smallest rejection region.

CHAPTER 3 PROPENSITY SCORES FOR KEY PERFORMANCE OBJECTIVES

The propensity score is defined as the conditional probability of assignment to a particular treatment given a set of observed covariates. The motivation for implementing propensity score methods is to transform the data to the probability scale and reduce variation among the employees. The propensity score matching allows the removal of selection bias between the treatment and control group. In other terms, the goal is to observe the effect of changes in key performance indicators on attrition rates. Since the experimental dataset is considered an observational dataset, we do not know that any differences in attrition rates will be solely due to employee responses to key performance indicators. However, there are other benchmark influential factors, such as demographics and education, that led employees with unfavorable KPI responses towards attrition and those with favorable responses towards retention.

There are two assumption with causality before we can implement propensity scores. They are the endogeneity and the ignorable treatment assignment assumptions. Suppose that there exists a binary treatment T, an outcome Y, and covariates X. The propensity score is defined as the conditional probability of treatment given background variables:

$$P(X) = Pr(T = 1|X = x). \tag{3.0.1}$$

Let $Y(0)$ and $Y(1)$ denote the potential outcomes under control and treatment, respectively. Then treatment assignment is (conditionally) unconfounded if potential outcomes are independent of treatment conditional on covariates X. This can be written as

$$Y(0), Y(1) \perp T(X), \tag{3.0.2}$$

where $\perp$ indicates statistical independence. If uncounfoundness holds, then

$$P(X) = Pr(T = 1|X = x). \tag{3.0.3}$$

In a two-group (case-control) experiment with random assignments, the probability of each individual in the sample to be assigned to the treatment is $P(Z = i|X) = 0.5$. In a quasi-experiment or observational study, the probability is unknown but it can be estimated from the data using a logistic regression model, where treatment assignment is regressed on the set of observed covariates. The PS then allows matching of the individuals in the case and control conditions with the same likelihood of receiving treatment. Propensity score matching employs a predicted probability of group membership (treatment vs. control group) based on observed predictors, usually obtained from logistic regression to create a counterfactual group. Thus, a pair of participants sharing a similar propensity score are seen as comparable, even though they may differ on values of specific covariates.

The matching technique used in this study is the nearest neighbor matching procedure. Near neighbor matches individuals from the case to participants in the control group based on distance. A participant (j) with propensity score P_j in the control sample (I_0) is a match for a participant (i) with propensity score P_i in the case group if the absolute difference between their propensity scores is the smallest.

$$C(P_i) = min||P_i - P_j||, j \in I_0 \tag{3.0.4}$$

The output in Table 3.1 below gives the count of matched cases for both sets of the key performance objective logistic equations after performing nearest neighbor one-to-one matching. Here we have 237 treated units and 1233 control units.

Table 3.1 Summary Statistics of Propensity Score Matches

	Control	Treated
All	1233	237
Matched	237	237
Unmatched	996	0
Discarded	0	0

Tables 3.2 and 3.3 provides the mean difference in the treated versus control cases for pre-

matching and post-matching.[1] The tables depict a stratified response variable to check the balance in the dataset among the employees who have exited and those who have remained. Note that the treated case is for those employees where attrition occurred and that the control case is for those who remain employed. After matching, nearly all mean differences have reduced greatly and have better balance in the dataset to proceed with further analysis. [2]

Table 3.2 Balance Statistics for Matched Culture KPO Pairs

Parameter		Means Treated	Means Control	Mean Difference
Job Satisfaction	Low	0.31 (0.28)	0.33 (0.18)	-0.02 (0.10)
	Medium	0.19 (0.19)	0.16 (0.19)	0.03 (0.00)
	High	0.31 (0.31)	0.33 (0.30)	-0.02 (0.01)
	Very High	0.22 (0.32)	0.23 (0.33)	-0.01 (-0.01)
Environment Satisfaction	Low	0.30 (0.30)	0.29 (0.17)	0.01 (0.13)
	Medium	0.18 (0.18)	0.17 (0.20)	0.01 (-0.02)
	Very High	0.25 (0.25)	0.25 (0.31)	0.00 (-0.06)
Education	< College	0.13 (0.13)	0.12 (0.11)	0.01 (0.02)
	College	0.19 (0.18)	0.18 (0.19)	0.00 (-0.01)
	Doctor	0.02 (0.02)	0.03 (0.04)	-0.01 (-0.02)
Department	Sales	0.39 (0.39)	0.43 (0.29)	-0.04 (0.10)
	R&D	0.56 (0.56)	0.54 (0.67)	0.03 (-0.13)
Marital Status	Single	0.51 (0.51)	0.51 (0.28)	0.00 (0.23)
	Married	0.35 (0.35)	0.38 (0.48)	-0.03 (-0.13)
Gender	Male	0.63 (0.63)	0.68 (0.59)	-0.05 (0.04)
Age		33.61 (33.61)	34.42 (37.56)	-0.81 (-3.94)

[1]The values in parenthesis are the post-matching means.

[2]More formal testing procedures exist; however, it is outside the goal of this paper.

Table 3.3 Balance Statistics for Matched Work Life Balance KPO Pairs

Parameter		Means Treated	Means Control	Mean Difference
Distance		10.63 (10.03)	11.11 (8.92)	-0.48 (1.11)
Work Life Balance	Bad	0.11 (0.05)	0.11 (0.11)	0.00 (-0.06)
	Better	0.54 (0.54)	0.56 (0.62)	-0.03 (-0.08)
	Good	0.24 (0.25)	0.23 (0.23)	0.02 (0.02)
	Best	0.11 (0.11)	0.11 (0.10)	0.01 (0.01)
Education	< College	0.13 (0.13)	0.15 (0.11)	-0.02 (0.02)
	College	0.19 (0.19)	0.20 (0.19)	-0.02 (0.00)
	Master	0.24 (0.25)	0.22 (0.28)	0.03 (-0.03)
	Doctor	0.02 (0.02)	0.02 (0.04)	0.00 (-0.02)
Department	Sales	0.39 (0.39)	0.41 (0.29)	-0.02 (0.10)
	R&D	0.56 (0.56)	0.52 (0.67)	0.04 (-0.11)
Marital Status	Single	0.51 (0.28)	0.47 (0.51)	0.04 (-0.23)
	Married	0.35 (0.35)	0.41 (0.48)	-0.05 (-0.13)
Gender	Male	0.63 (0.63)	0.60 (0.59)	0.03 (-0.04)
Age		33.61 (33.61)	33.71 (37.56)	-0.10 (-3.95)

CHAPTER 4 OUTCOME ANALYSIS

We will now conduct an outcome analysis on the matched propensity scores to test which key performance indicator has the highest probability of decreasing attrition when employee satisfaction responses are increased. Note that the most significant contrasts found in Chapter 2 for all key performance indicators is the difference between the most favorable and least favorable responses. The outcome analysis will perform a paired t-test on the matched propensity scores between these two response levels for each key performance indicator. It is important to implement the paired t-test on the difference in propensity scores as opposed to an unpaired t-test due to the strong positive correlation among the matched responses. [1]

The output in Table 4.1-4.3 gives the paired t-test results on the difference in the outcome of the matched pairs. Before performing the paired t-test, two subsets from the matched propensity score data, one for most favorable responses and another for least favorable responses, were created. The number of observations then decreased for each key performance indicator; 18 observations for job satisfaction, 24 observations for environment satisfaction, and 7 observations for work life balance. The goal is to test the hypothesis of no difference in means on the matched data between these two types of employees and determine (1) which key performance indicator has the overall highest probability of reducing attrition and (2) which increases in satisfaction levels lead to a significant difference in attrition rates.

From the results we see the key performance indicator work life balance has the only significant difference in means between the least favorable and most favorable satisfaction levels at an α level of 0.05. The interpretation here is that increasing work life balance satisfaction from bad to best will have a significant effect on reducing attrition rates, wheres the same cannot be stated for the culture key performance objective. We can rank the importance of the key performance indicators on the probability of attrition by comparing the means. The figure and output values show that the key performance indicator associated with the highest rates of attrition, after accounting for

[1] See Appendix C for the output of an unpaired t-test.

difference in individual characteristics, is work life balance followed by environment satisfaction and job satisfaction, respectively.

Table 4.1 Job Satisfaction Paired t-test for Significant Contrasts

Job Satisfaction		
Mean	0.194762	0.194491
Variance	0.008389	0.008205
Observations	18	18
Pearson Correlation	0.999752	
df	17	
t-statistic	0.507669	
P(T ≤ t) one-tail	0.309104	
P(T ≤ t) two-tail	0.618209	

Table 4.2 Environment Satisfaction Paired t-test for Significant Contrasts

Environment Satisfaction		
Mean	0.278414	0.27793
Variance	0.020451	0.020663
Observations	24	24
Pearson Correlation	0.991448	
df	23	
t-statistic	0.126497	
P(T ≤ t) one-tail	0.450219	
P(T ≤ t) two-tail	0.900438	

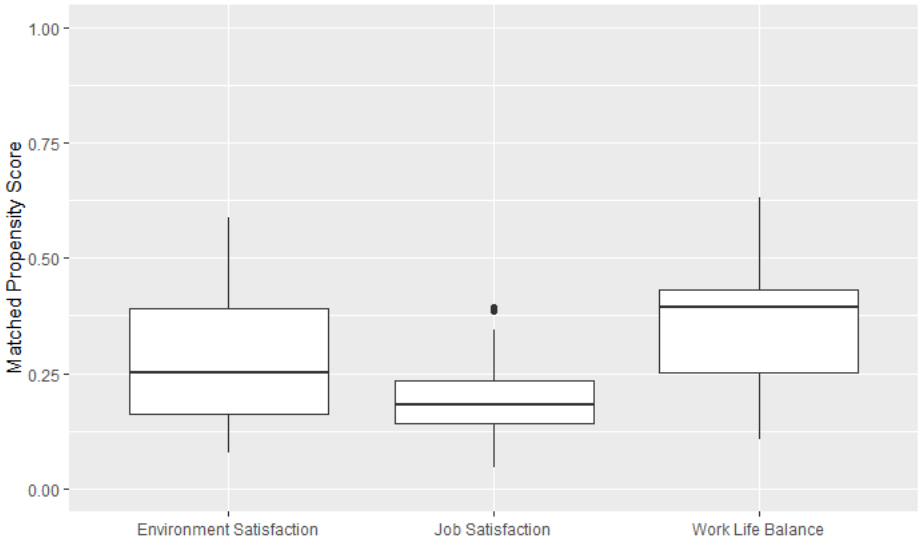
Table 4.3 Work Life Balance Paired t-test for Significant Contrasts

Work Life Balance		
Mean	0.372043524	0.358852
Variance	0.029029886	0.024905
Observations	7	7
Pearson Correlation	0.998426842	
df	6	
t-statistic	2.241037726	
P(T ≤ t) one-tail	0.033126231	
P(T ≤ t) two-tail	0.066252463	

The boxplot below illustrates these differences in probability rates of attrition. The key performance indicator with the highest mean rate of attrition between satisfaction favorability responses

can be viewed as the KPI most associated with attrition. In other words, dedicating resources to improving work life balance satisfaction has the highest probability of successfully reducing the rate of attrition as compared to the other key performance indicators despite insigifcant differences in means.

Figure 4.1 Boxplot of Propensity Score Matches



CHAPTER 5 CONCLUSION

This thesis sets forth an analysis for business leaders attempting to understand their organizations human capital to retain top talent. Key performance indicators allow any organization to better understand and manage strategic business initiatves, but determing which key performance indicator to focus time and resources towards may pose challenges. With regards to the experimental dataset in this report, we have contrusted a method to rank the order of importance of key performance indicators for reducing the rate of employee attrition. From the previous analyses, it is clear that the KPI work life balance has the highest probability of successfully reducing attrition when resources are focused on increasing work life balance satisfciation from least favorale to most favorable. Additionally, despite the fact the changes to the key performance objectives of culture are not statisitcally significant in reducing attrition rates when increasing favorability responses, we are still able to define a strategic plan to retain employees. As shown in Chapter 4, any strategic plan with the goal of reducing attrition based upon relevant key performance indicators should focus first on work life balance, then on environment satisfaction and job satisfaction, respectively. Thus, the methodology proposed can successfully utilize organizational observational data along with survey results to make data-driven decisions on human capital. Future research will be able to collect longitudnal data and conduct an approriate analysis over time to gain more insight into the effects of bussiness decisions on human capital and attrition rates of its employeees.

BIBLIOGRAPHY

[1] He, H., Wu, P. and Chen, D. (2016) *Statistical Causal Inferences and Their Applications in Public Health Research.* Springer International.

[2] Rosenbaum, P., and Rubin, D. (1983). *The Central Role of the Propensity Score in Observational Studies for Causal Effects.* Biometrika, 70(1), 41-55. doi:10.2307/2335942

APPENDIX A SELECTED SAS PROGRAMS

```
proc logistic data=workforce;
   class jobsatisfaction environmentsatisfaction education department gender maritalstatus;
   model attrition= jobsatisfaction environmentsatisfaction education department gender maritalstatus age
      /selection=forward expb;
run;
```

```
proc logistic data=workforce plots=EFFECT plots=ROC;
   class jobsatisfaction environmentsatisfaction education department maritalstatus gender;
   model attrition= jobsatisfaction environmentsatisfaction education department maritalstatus gender age
      / outroc=rocout;
   output out=estimated predicted=estprob l=lower95 u=upper95;
run;
```

```
proc logistic data=workforce plots(only)=(oddsratio(range=clip));
   class jobsatisfaction environmentsatisfaction education department maritalstatus gender;
   model attrition= jobsatisfaction environmentsatisfaction education department maritalstatus gender age;
   oddsratio jobsatisfaction;
   oddsratio environmentsatisfaction;
   oddsratio education;
   oddsratio department;
   oddsratio maritalstatus;
   oddsratio age;
   contrast 'Job Satisfaction'
      jobsatisfaction 0 -1 0 1,
      jobsatisfaction 1 -1 0 0,
      jobsatisfaction 0.5 -0.5 0.5 -0.5 / estimate=exp;
   contrast 'Environment Satisfaction'
      environmentsatisfaction 0 -1 0 1,
     environmentsatisfaction 1 -1 0 0,
     environmentsatisfaction 0.5 -0.5 0.5 -0.5 / estimate=exp;
   effectplot / at(environmentsatisfaction=all) noobs;
   effectplot slicefit(sliceby=environmentsatisfaction plotby=jobsatisfaction) / noobs;
run;
```

```
proc  logistic data=workforce plots=EFFECT plots=ROC;
   class jobsatisfaction environmentsatisfaction;
   model attrition= jobsatisfaction environmentsatisfaction / outroc = rocout;
   output out=estimated predicted=estprob l=lower95 u=upper95;
run;
```

```
proc logistic data=workforce;
   class worklifebalance education department gender maritalstatus;
   model attrition= distance worklifebalance education department gender maritalstatus age
      /selection=forward expb;
run;
```

```
proc logistic data=workforce plots=EFFECT plots=ROC;
   class worklifebalance education department maritalstatus gender;
   model attrition= worklifebalance distance education department maritalstatus gender age
      / outroc=rocout;
   output out=estimated predicted=estprob l=lower95 u=upper95;
run;
```

```
proc logistic data=workforce plots(only)=(oddsratio(range=clip));
   class worklifebalance education department maritalstatus gender;
   model attrition= worklifebalance distance education department maritalstatus gender age;
   oddsratio worklifebalance;
   oddsratio education;
   oddsratio department;
   oddsratio maritalstatus;
   oddsratio age;
   oddsratio gender;
   contrast 'Work Life Balance'
      worklifebalance -1 1 0 0,
      worklifebalance 0 1 -1 0,
      worklifebalance -0.5 0.5 -0.5 0.5 / estimate=exp;
   effectplot / at(worklifebalance=all) noobs;
run;
```

```
proc  logistic data=workforce plots=EFFECT plots=ROC;
   class worklifebalance ;
   model attrition= worklifebalance / outroc = rocout;
   output out=estimated predicted=estprob l=lower95 u=upper95;
run;
```

APPENDIX B SELECTED R PROGRAMS

```
library(dplyr)
library(MatchIt)
library(ggplot2)
library(WhatIf)
library(knitr)

workforce %>%
  group_by(attrition) %>%
  summarise(n_employees = n())

workforce %>%
  group_by(worklifebalance) %>%
  summarise(n_employees = n(),
            mean_attriton = mean(attrition))

workforce %>%
  group_by(jobsatisfaction) %>%
  summarise(n_employees = n(),
            mean_attriton = mean(attrition))

workforce %>%
  group_by(environmentsatisfaction) %>%
  summarise(n_employees = n(),
            mean_attriton = mean(attrition))




workforce_contrastI %>%
  group_by(attrition) %>%
  select(c('culturejob', 'cultureenvironment', 'worklife')) %>%
  summarise_all(funs(mean(., na.rm = T)))

workforce_culture <- c(37, 38, 39)
lapply(workforce_culture, function(v) {
   t.test(workforce_contrastI[, v] ~ workforce_contrastI [, 'attrition'])




m_ps_culture <- glm(attrition ~ jobsatisfaction + environmentsatisfaction + education +
   department + maritalstatus + gender + age, family = binomial("logit"), data = workforce)

prs_df_culture <- data.frame(pr_score = predict(m_ps_culture, type = "response"),
   attrition = m_ps_culture$model$attrition)

match_culture <- matchit(m_ps_culture, method="nearest", data=workforce)

pairs_culture <- matrix(c(as.integer(row.names(match_culture$match.matrix)),
   as.integer(match_culture$match.matrix[,1])), ncol=2)

n=NROW(pairs_culture)
culture<- matrix(0:0, n, 6)
for(i in 1:n) {
  culture[i,1]<-workforce[pairs_culture[i,1],17] # first matched pair job
  culture[i,2]<-workforce[pairs_culture[i,2],17] #second match pair job
  culture[i,3]<-m_ps_culture$fitted.values[pairs_culture[i,1]] #first propensity score
  culture[i,4]<-m_ps_culture$fitted.values[pairs_culture[i,2]] #second propensity score
  culture[i,5]<-workforce[pairs_culture[i,1],32] # first matched pair job
  culture[i,6]<-workforce[pairs_culture[i,2],32] #second match pair job
```

```
m_ps_balance <- glm(attrition ~ worklifebalance + distance + education + department +
   maritalstatus + gender + age, family = binomial("logit"), data = workforce)

prs_df_balance <- data.frame(pr_score = predict(m_ps_balance, type = "response"),
   attrition = m_ps_balance$model$attrition)

match_balance <- matchit(m_ps_balance, method="nearest", data=workforce)

pairs_balance <- matrix(c(as.integer(row.names(match_balance$match.matrix)),
   as.integer(match_balance$match.matrix[,1])), ncol=2)




n=NROW(pairs_balance)
balance<- matrix(0:0, n, 6)
for(i in 1:n) {
  balance[i,1]<-workforce[pairs_balance[i,1],34] # first matched pair job
  balance[i,2]<-workforce[pairs_balance[i,2],34] #second match pair job
  balance[i,3]<-m_ps_balance$fitted.values[pairs_balance[i,1]] #first propensity score
  balance[i,4]<-m_ps_balance$fitted.values[pairs_balance[i,2]] #second propensity score
}
```

APPENDIX C ADDITIONAL OUTPUT FOR KEY PERFORMANCE OBJECTIVES

Culture Logistic Equations AIC

Model Fit Statistics	
AIC	1203.002

Balance Logisitc Equations AIC

Model Fit Statistics	
AIC	1214.325

Logisitc Regression Analysis of Culture Maximum Likelihood Estimates

Parameter		DF	Estimate	Std. Error	Chi-Sq	Pr>ChiSq
Intercept		1	-0.0513	0.3574	0.0206	0.8859
Job Satisfaction	Low	1	-0.4636	0.1297	12.770	0.0004
	Medium	1	-0.0126	0.1420	0.0079	0.9293
	High	1	-0.0078	0.1215	0.0041	0.9489
Environment Satisfaction	Low	1	-0.5986	0.1276	22.010	<.0001
	Medium	1	0.1190	0.1435	0.6884	0.4067
	High	1	0.2274	0.1264	3.2352	0.0721
Education	< College	1	-0.0391	0.2061	0.0359	0.8497
	College	1	-0.0843	0.1781	0.2243	0.6358
	Bachelor	1	-0.1285	0.1488	0.7456	0.3879
	Doctor	1	0.2987	0.3948	0.5724	0.4493
Department	HR	1	-0.2222	0.2317	0.9196	0.3376
	R&D	1	0.3571	0.1368	6.8130	0.0090
Marital Status	Divorced	1	0.4512	0.1371	10.823	0.0010
	Married	1	0.1887	0.1086	3.0162	0.0824
Gender	Female	1	0.1033	0.0775	1.7736	0.1829
Age		1	0.0472	0.0092	26.627	<.0001

Logisitc Regression Analysis of Work Life Balance Maximum Likelihood Estimates

Parameter		DF	Estimate	Std. Error	Chi-Sq	Pr>ChiSq
Intercept		1	0.0300	0.3709	0.0066	0.9355
Distance		1	-0.0287	0.0088	10.687	0.0011
Work Life Balance	Bad	1	-0.7245	0.2049	12.503	0.0004
	Better	1	0.4258	0.1191	12.787	0.0003
	Best	1	0.1508	0.1836	0.6752	0.4112
Education	< College	1	-0.0244	0.2039	0.0143	0.9047
	College	1	-0.0573	0.1766	0.1052	0.7457
	Bachelor	1	-0.1143	0.1472	0.6035	0.4372
	Doctor	1	0.2437	0.3911	0.3884	0.5332
Department	HR	1	-0.2242	0.2306	0.9451	0.3310
	R&D	1	0.3660	0.1362	7.2156	0.0072
Marital Status	Divorced	1	0.4288	0.1366	9.8605	0.0017
	Married	1	0.2113	0.1083	3.8091	0.0510
Gender	Female	1	0.0990	0.0770	1.6539	0.1984
Age		1	0.0465	0.0091	25.960	<.0001

Contrast Confidence Intervals

Contrast		Confidence Limits	
Job Satisfaction	Very High // Low	1.2329	2.0502
	High // Low	1.0652	2.3362
	Upper // Lower	0.9773	1.5940
Environment Satisfaction	Very High // Low	1.4169	2.3365
	High // Low	1.5318	3.4060
	Upper // Lower	1.2556	2.0492

Balance Contrast Confidence Intervals

Contrast		Confidence Limits	
Work Life Balance	Best // Bad	1.2362	4.6587
	Best // Better	0.4722	1.2219
	Upper // Lower	0.8914	1.7586

Job Satisfaction t-test for Significant Contrasts

Job Satisfaction		
t-statistic	4.1318	
df	159.98	
p-value	5.789e-05	
Confidence Interval	0.107	0.304

Environment Satisfaction t-test for Significant Contrasts

Environment Satisfaction		
t-statistic	4.0024	
df	187.75	
p-value	9.02e-05	
Confidence Interval	0.097	0.285

Work Life Balance t-test for Significant Contrasts

Work Life Balance		
t-statistic	2.2706	
df	77.169	
p-value	0.02596	
Confidence Interval	0.0218	0.332