*Statistical Analysis Plan for Firm 3*

# <version 1>

<Sonkashi Sharma, Brooke Woods, Ugochi Madumere>

<July 20, 2020>

**C contents**

## Introduction

“Global Talent Monitor’s report on workforce activity in 2018 shows that the lack of future career development remains a key driver of employee attrition — cited by 40% of departing employees as a dissatisfying factor in their job. At the same time, 28% of employees are actively seeking a job and 42% are passively open to new opportunities (Gartner 2018)”. “Employee attrition costs large organizations millions of dollars each year and the loss of a particularly conscientious employee can be debilitating, not just to culture and morale, but to employee productivity(Gartner 2018)”.This study intends to understand the factors which contribute to employee attrition in an organization. The study would contribute towards providing an insight for organizations which could eventually reduce their financial losses caused due to employee attrition.

IBM, also known as Business Process Manager on cloud, is known for producing and selling computer hardware, middleware and software, and providing hosts and consulting services to businesses. It is also well known for its research organization and its subscription services that provide clients a full lifecycle business process management (BPM) environment which includes the development, test, and production with tooling and run times for the process design, process execution, process monitoring,, and optimization. IBS offers the visualization and management of business processes, with low startup costs and a high return on investment for businesses. ¹

For this project, we will look specifically at the onboarding vs the recruiting process to map out the process for the employees of IBS. We will look at the univariate and multivariate analytics of employees and the industry in general to find variables which contribute to high employee attrition rates. We will then use software like SAS Enterprise Miner, SAS Studio and SPSS to run a binary logistic regression on our model which will target the variables contributing to high attrition rates and therefore, the variables which influence employee behavior. We will compare the activities of the employment process using the metrics of design, management, and improvement.

Predictive analytics can be to better help organizations understand and design interventions that will be most effective in reducing unwanted attrition.  Financial considerations aside, businesses are better off when they can retain good employees and the organizational knowledge they possess. Our findings will be oriented to helping organizations understand what is most important to their employees, with the goal of making improvements to increase employee engagement and productivity and reduce unwanted attrition. The comparison of engagement survey data to termination data can reveal areas of the employee experience in need of improvement. Ultimately, employee attrition can design an employee retention model that will work even if attrition is not expected to be a big issue soon. The model will still help the organization determine if it is experiencing the right kind of attrition. ⁴

***Data Source***

For this project, we will analyze a dataset called “IMB HR Analytics Employee Attrition & Performance” which is a fictional dataset from IBM data scientists downloaded from the data source Kaggle. This dataset is composed of 1479 variables and 35 columns, containing variables such as age, travel frequency, daily rate, distance from home, education, employee count, department, etc. It is available for download at [https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/data#](https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/data). ⁴

*Exit surveys are another potential data source that can provide richer information. Comparing responses on exit surveys to employees’ engagement survey responses can reveal how the employees’ perceptions changed over time. Correlating exit and engagement survey data can yield additional capability to predict attrition risk.*

## Analysis Objectives

* *Understanding the factors which will influence the employer and lead them to employee attrition*
* *In the comparison, looking into parameters like model fit, R squared comparison, multicollinearity (VIF, tolerance), and P values significance.*

## Analysis Description

*-Logistic regression* must be used since the target is a binary variable. Logistic regression is used to predict the odds of being a case based on the values on the independent variables (predictor variables). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non case.

So, in our project we test robustness using the relative insensitivity of the statistical test to violations of the underlying internal assumptions:

* 1. *Independence of the residuals* means that the residuals are not autocorrelated. The assumption of independence can be tested using scatter plots: plot of residuals versus case number. The points should be symmetrically distributed around a horizontal line. The Durbin-Watson statistic can also be used to test for independence of the residuals. If the Durbin-Watson statistic is close to 2, there is no autocorrelation. Violation of this assumption can create bias in the significance tests and confidence intervals. ⁶

*Remedy: One of the examples of violation occurs when we have time relevant or longitudinal data. The best way to analyze such data is General Estimation Equation (GEE).*

* 1. *Normality of the residuals* means that the residuals follow a normal distribution. Assumes 99% of values will fall within 3 standard deviations of the mean. The assumption of normality can be tested using plots: Q-Q plots; histograms. When plotting Q-Q plot of observed versus expected values, the points should be symmetrically distributed around a diagonal line with z values greater to or less than 3 being considered outliers. When plotting the histogram of residuals, the shape should be of a normal distribution. Observations will be arranged in increasing magnitude and plotted against normal expected distribution. If the test is non-statistically significant at 5%, there is no evidence of violation of the normality assumption. Violation of these assumptions can lead to bias in the estimation of coefficients and standard errors especially when the sample size is small. Nonnormality does not lead to severe problems in the interpretation when the sample size is large. Our sample size is considered large since it is >30. ⁶

*Remedy: Data transformation, and particularly the Box-Cox power transformation, is one of the remedial actions that may help to make data normal.*

* 1. *Collinearity of residuals* exists if there is an approximate linear relationship (i.e., shared variance) among some of IVs in the data.

*Problem with collinearity:* Collinearities inflate the variances of the regression coefficients. This could have the following consequences: Parameter estimates that fluctuate dramatically with negligible changes in the sample; parameter estimates with signs that are "wrong" in terms of theoretical considerations; theoretically "important" variables with insignificant coefficients; the inability to determine the relative importance of collinear variables. ⁶

*What can multicollinearity do?* It produces large SE of Bs which results in coefficients being non-significant; it produces bizarre β estimates (e.g., wrong direction); Removal or addition of one IV results in enormous change to the models.

*When are multicollinearity problems certain*? Tolerance ≤ .1 (or .2). [VIF ≥ 10 (or 5)]; even if tolerance ≥

.2 (or VIF ≤ 5), multicollinearity could be problematic when: There is a bivariate correlation of .7 or more between two IVs; the bivariate correlation between two IVs are greater than either of IV’s correlation with DV. ⁶

*Remedies for multicollinearity*: Omit the redundant IV [or aggregate the two similar ones] (Hair et al. 2010); Increase the sample size (Hair et al. 2010); Transform the raw-data X to create a new, orthogonal matrix (Mason and Perreault 1991); Mean center or Scale center raw X.

## Statistical methodology/procedures

Before the implementation of any algorithm, we must split the data into training set (70%), validation set (20%) and testing set (10%) using SAS EM’s partition node.

In the logit model the log odds of the outcome is modeled as a linear combination of predictor variables. We will name the node SAS EM node Logit. With Logistic regression, the training and predicting speed is relatively very fast.

*Using Multivariate analytics to detect the outliers*

Some outliers could be detected from the Mahala Nobis distance which detects outliers of any kind. The Mahala Nobis distance should be divided by degrees of freedom (1 subtracted from the variables which would be entered as IV’s).

For multivariate outliers, mahala Nobis distance is evaluated on a chi squared (x2) statistics with degrees of freedom=# of variables in the analysis. Testing the significance of the model can be done using chi square significant values with p values <.05 accepted and considered significant.⁶

All hypothesis testing will be done at 5% (95% confidence interval) with a 2-sided test with a significance level of <.05 unless otherwise specified for categorical parameters.

P values are rounded using 3 decimal places and values <.001 will be reported as <.001 Normality assumption can be tested using Kolmogorov-Smirnov and Shapiro-Wilk. ⁶

## Handling of missing data/outliers

An initial cleaning of this dataset will be needed before an analysis can be performed. The dataset will be examined for missing data patterns. We will estimate missing values using a logistic approach and later use them during the main analysis. We will find univariate (extreme values of one variable) and multivariate (unusual combinations of 2 or more z scores) outliers. Outliers can be detected by visually inspecting the data with frequency distribution/histograms of the residuals and with standardization. Standardization is applied to transform the data to z scores. Z scores are tested using the normality assumption, and outliers are detected by values with Z scores falling greater than 3 or less than -3.

For outlier detection we are going to perform the following analysis:

**Graph Examination & Outlier detections:**

*Outlier detection:*

## Apply a Standardized score to identify outliers for the continuous variables.

Applying a standardized score to identify outliers transforms the skewness of our model for the continuous variables falling within the range of -3 to 3. Such variables would not be considered as outliers in this business process. We would also look at the Box plot for these variables also to check for outliers. ⁶

## Apply box-whisker plot to identify the outliers for the continuous variables

We would analyze the values shown in the dataset that approach the threshold -3 and 3. Since the dataset is so small, it is possible that the values above and below 2 could be considered as outliers. A graphical method can be used to examine outliers in the boxplot. The lines extend to the outliers in the

box-whiskers plot. ⁶

## Apply scatter plot matrix to identify the outliers for the continuous variables.

We will choose individual bivariate scatterplots to spot the outliers and choose the Z scores instead of using original scores. This could give a clearer picture of the observations. ⁶

OUR FIRM'S ETHICAL CODE

We pledge in writing to abide by the American Statistical Association's (ASA) and INFORMS' Codes of Ethics. Our adherence to these Codes signifies voluntary assumption of self-discipline. As the professional associations for our firm in the United States, the ASA and INFORMS requires adherence to their Codes of Ethics as a condition of membership. The standards of conduct set forth in these Codes provide basic principles in the ethical

practice of data analysis consulting. The purpose of these Codes is to help us maintain our professionalism and adhere to high ethical standards in the conduct of providing services to clients and in our dealings

with our colleagues and the public. Our individual judgment requires we apply these principles. We are liable to disciplinary action under the ASA's and INFORMS' Rules of Procedure for Enforcement of this Code if our conduct is found by the ASA's or INFORMS' respective Ethics Committees to be in violation of their respective Codes or to bring discredit to the profession or to ASA and INFORMS .

# Our Commitment to Our Clients

* 1. We will serve our clients with integrity, competence, independence, objectivity, and professionalism.
  2. We will mutually establish with our clients realistic expectations of the benefits and results of our services.
  3. We will only accept assignments for which we possess the requisite experience and competence to perform and will only assign staff or engage colleagues with the knowledge and expertise needed to serve our clients effectively.
  4. Before accepting any engagement, we will ensure that we have worked with our clients to establish a mutual understanding of the objectives, scope, work plan, and fee arrangements.
  5. We will treat appropriately all confidential client information that is not public knowledge, take reasonable steps to prevent it from access by unauthorized people, and will not take advantage of proprietary or privileged

information, either for use by ourselves, the client's firm, or another client, without the client's permission.

* 1. We will avoid conflicts of interest or the appearance of such and will immediately disclose to the client circumstances or interests that we believe may influence my judgment or objectivity.
  2. We will offer to withdraw from a consulting assignment when we believe my objectivity or integrity may be impaired.
  3. We will refrain from inviting an employee of an active or inactive client to consider alternative employment without prior discussion with the client.

# Our Commitment to Fiscal Integrity

* 1. We will agree in advance with a client on the basis for fees and expenses and will charge fees that are reasonable and commensurate with the services delivered and the responsibility accepted.
  2. We will not accept commissions, remuneration, or other benefits from a third party in connection with the recommendations to a client without that client's prior knowledge and consent, and will disclose in advance any financial interests in goods or services that form part of such recommendations.

# Our Commitment to the Public and the Profession

* 1. If within the scope of my engagement, we will report to appropriate authorities within or external to the client organization any occurrences of malfeasance, dangerous behavior, or illegal activities.
  2. We will respect the rights of consulting colleagues and consulting firms and will not use their proprietary information or methodologies without permission.
  3. We will represent the profession with integrity and professionalism in my relations with our clients, colleagues, and the general public.
  4. We will not advertise our services in a deceptive manner nor misrepresent or denigrate individual consulting practitioners, consulting firms, or the consulting profession.
  5. If we perceive a violation of the Code, we will report it to the APA and INFORMS and will promote adherence to the Code by other member consultants working on our behalf.

# Sonakshi Sharma Brooke Woods Ugochi Madumere

*Works Cited*

*“Gartner Survey Shows 29 Percent of Employees Witnessed At Least One Compliance Violation In*

*The Last Two Years.” Gartner,*

[*w ww.gartner.com/en/newsroom/press-releases/2018-08-02-gartner-survey-shows-29-pe*](http://www.gartner.com/en/newsroom/press-releases/2018-08-02-gartner-survey-shows-29-pe)

*rcent-of-employees-witnessed-at-least-one-compliance-violation-in-the-last-two-years.*¹

“Here's How IBM Predicts 95% of Its Turnover Using Data.” *LinkedIn Talent Blog*,

# business.linkedin.com/talent-solutions/blog/artificial-intelligence/2019/IBM-predicts-95-\ percent-of-turnover-using-AI-and-data.²

“Lack of Career Development Drives Employee Attrition.” *Smarter With*

*Gartner*,[www.gartner.com/smarterwithgartner/lack-of-career-development-drives-emplo](http://www.gartner.com/smarterwithgartner/lack-of-career-development-drives-employ) [y](http://www.gartner.com/smarterwithgartner/lack-of-career-development-drives-employ)

ee-attrition/. ³

Pavansubhash. “IBM HR Analytics Employee Attrition & Performance.” *Kaggle*, 31 Mar.

[2017, www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.](http://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset)⁴

Rohan’s Four - Rohan Jain, Ali Shahid. *IBM HR Analytics Employee Attrition & Performance*,inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/Januar y2018FBL/IBM\_Attrition\_VSS.html⁵

Zaal, T.M.E, and Steve Newton. *Integrated Design and Engineering: as a Business Improvement*

*Process*. Maj Engineering Publishing, 2014. ⁶