**ISOM 631 – FINAL PROJECT**

**IBM - EMPLOYEE ATTRITION BASED ON HR DATA**

**-NISHI TYAGI**

/\*IMPORT THE DATA\*/

**PROC** **IMPORT** OUT=WORK.FINAL\_PROJ\_NT (DROP=EmployeeCount EmployeeNumber Over18 StandardHours)

DATAFILE='\\adm.suffolk.edu\uem\STD-RedirectedFolders\ntyagi\Documents\ISOM 631\FINAL PROJECT\attrition.xlsx'

DBMS=XLSX REPLACE;

SHEET='WA\_Fn-UseC\_-HR-Employee-Attriti';

GETNAMES=yes;

**RUN**;

/\*Data Dictionary for some categorical data\*/

/\*

Education

1 'Below College'

2 'College'

3 'Bachelor'

4 'Master'

5 'Doctor'

EnvironmentSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

JobInvolvement

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

JobSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

PerformanceRating

1 'Low'

2 'Good'

3 'Excellent'

4 'Outstanding'

RelationshipSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

WorkLifeBalance

1 'Bad'

2 'Good'

3 'Better'

4 'Best'

\*/

/\*We drop the attributes Over18, EmployeeCount and StandardHours, because they all carry the same value for each row

We drop the attribute EmployeeNumber as it has unique values for each observation.\*/

/\*FIND CORRELATION\*/

**PROC** **CORR** DATA=WORK.FINAL\_PROJ\_NT;

VAR Age DailyRate DistanceFromHome Education EnvironmentSatisfaction HourlyRate

JobInvolvement JobLevel JobSatisfaction MonthlyIncome MonthlyRate NumCompaniesWorked PercentSalaryHike PerformanceRating

RelationshipSatisfaction StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance

YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager;

WITH Age DailyRate DistanceFromHome Education EnvironmentSatisfaction HourlyRate

JobInvolvement JobLevel JobSatisfaction MonthlyIncome MonthlyRate NumCompaniesWorked PercentSalaryHike PerformanceRating

RelationshipSatisfaction StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance

YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager;

**RUN**;

/\*Job level and Monthly income are highly correlated. Hence, I will exclude Job level from my data set. Hence, the new data will be as below\*/

**DATA** WORK.FINAL\_PROJ\_NT2(DROP=JobLevel);

SET WORK.FINAL\_PROJ\_NT;

**RUN**;

/\*DERIVED COLUMNS\*/

/\*I will be a computing a column - JobHop, which will give us information about how many jobs a person has switched. This can be a reason for Employee attrition.

The formula is given by - JobHop = TotalWorkingYears/NumCompaniesWorked\*/

**DATA** WORK.FINAL\_PROJ\_NT3;

SET FINAL\_PROJ\_NT2;

JobHop = TotalWorkingYears/NumCompaniesWorked;

JobHop = round(JobHop, **2**);

**RUN**;

/\*Check missing values\*/

**PROC** **MEANS** N MIN MAX SKEWNESS NMISS MEAN DATA=WORK.FINAL\_PROJ\_NT3;

var Age DailyRate DistanceFromHome HourlyRate

JobHop MonthlyIncome MonthlyRate PercentSalaryHike TotalWorkingYears

YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager;

**RUN**;

/\*Handling missing values in JobHop as they mean that the person has never switched. Replace them with 0\*/

**PROC** **STDIZE** DATA=WORK.FINAL\_PROJ\_NT3

OUT=WORK.FINAL\_PROJ\_NT\_NO\_MISS

REPONLY MISSING=**0**;

**RUN**;

/\*Check for skewness and outliers using univariate and histogram\*/

**PROC** **MEANS** N MIN MAX SKEWNESS NMISS DATA=WORK.FINAL\_PROJ\_NT\_NO\_MISS;

var Age DailyRate DistanceFromHome HourlyRate

JobHop MonthlyIncome MonthlyRate PercentSalaryHike TotalWorkingYears

YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager;

**RUN**;

/\*Skewness is present for the variables - JobHop, MonthlyIncome,

We handle the skewness using log transform\*/

**DATA** WORK.FINAL\_PROJ\_NT\_NO\_SKEW;

SET WORK.FINAL\_PROJ\_NT\_NO\_MISS;

LogJobHop= log(JobHop+**1**);

LogMonthlyIncome= log(MonthlyIncome+**1**);

DROP JobHop MonthlyIncome;

**RUN**;

/\*Below code to check skewness:\*/

**PROC** **MEANS** N MIN MAX SKEWNESS DATA=WORK.FINAL\_PROJ\_NT\_NO\_SKEW;

var Age DailyRate DistanceFromHome HourlyRate

LogJobHop LogMonthlyIncome MonthlyRate PercentSalaryHike TotalWorkingYears

YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager;

**RUN**;

/\*We see that the skewness has reduced after taking the log transformation\*/

/\*We can confirm the following with the histograms generated through below code:\*/

**PROC** **UNIVARIATE** DATA=WORK.FINAL\_PROJ\_NT\_NO\_SKEW;

Histogram Age / midpoints= **10** to **70** by **10**;

Histogram DailyRate / midpoints= **100** to **1600** by **100**;

Histogram DistanceFromHome / midpoints= **0** to **30** by **5**;

Histogram HourlyRate / midpoints= **10** to **100** by **10**;

Histogram LogJobHop / midpoints= **0** to **50** by **5**;

Histogram LogMonthlyIncome / midpoints= **1000** to **20000** by **1000**;

Histogram MonthlyRate / midpoints= **2000** to **28000** by **1000**;

Histogram PercentSalaryHike / midpoints= **10** to **30** by **5**;

Histogram TotalWorkingYears / midpoints= **0** to **40** by **5**;

Histogram YearsAtCompany / midpoints= **0** to **50** by **5**;

Histogram YearsInCurrentRole / midpoints= **0** to **20** by **1**;

Histogram YearsSinceLastPromotion / midpoints= **0** to **20** by **1**;

Histogram YearsWithCurrManager / midpoints= **0** to **20** by **1**;

**RUN**;

/\*Attrition has to be converted to binary for my logistic regression model\*/

/\*Employee attrition Yes = 1

and no =0\*/

**DATA** WORK.FINAL\_PROJ\_NT\_TaregtVariable;

SET FINAL\_PROJ\_NT\_NO\_SKEW(RENAME=(Attrition=Attrition1));

IF Attrition1 = 'Yes' THEN Attrition1 = **1**;

ELSE Attrition1 = **0**;

IF PerformanceRating = **3** THEN PerformanceRating1 = 'High';

ELSE PerformanceRating1 = 'Very High';

DROP Attrition;

**RUN**;

/\*Using PROC FREQ for the categorical data

Checking the frequency and the percent of data for each category\*/

**PROC** **FREQ** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable;

TABLE Attrition1 BusinessTravel Department Education EducationField

EnvironmentSatisfaction Gender JobInvolvement JobRole JobSatisfaction

MaritalStatus NumCompaniesWorked OverTime PerformanceRating RelationshipSatisfaction

WorkLifeBalance;

**RUN**;

/\* We can see that the data is more populated for employees with BusinessTravel - Rarely as compared to the other two categories.

The percentage of employees in Research & Development is more as compared to Human Resources and Sales.

Similarly, Male employees are more as compared as Female employees.

More Employees have high JobSatisfaction, EnvironmentSatisfaction and JobInvolvement.

There is a high frequency of Married employees as compared to Single and Divorced.\*/

/\*Lets test some hypothesis\*/

/\*1\*/

/\*H0: There is no difference in the daily rate of males and females\*/

**PROC** **TTest** Data= WORK.FINAL\_PROJ\_NT\_TaregtVariable sides=**2** alpha=**0.05** h0=**0**;

class Gender;

var DailyRate;

**RUN**;

/\*As the p-value is greater than 0.05, we accept the null hypothesis that there is no difference in the daily rate of males and females\*/

/\*2\*/

/\*H0: The salary hike is indifferent of performance rating\*/

**PROC** **TTest** Data= WORK.FINAL\_PROJ\_NT\_TaregtVariable sides=**2** alpha=**0.05** h0=**0**;

class PerformanceRating1;

var PercentSalaryHike;

**RUN**;

/\*As the p-value is less than 0.05, we reject the null hypothesis and conclude

that the PerformanceRating effects the Salary Hike of an employee.\*/

/\*3\*/

/\*H0: There is no difference in the salary hike for male and female\*/

**PROC** **TTest** Data= WORK.FINAL\_PROJ\_NT\_TaregtVariable sides=**2** alpha=**0.05** h0=**0**;

class Gender;

var PercentSalaryHike;

**RUN**;

/\*As the p-value is greater than 0.05, we accept the null hypothesis and conclude

that there is no difference in the salary hike for male and female.\*/

/\*4\*/

/\*H0: The mean difference in YearsAtCompany and YearsInCurrentRole is 0\*/

**PROC** **TTest** Data=WORK.FINAL\_PROJ\_NT\_TaregtVariable sides=**2** alpha=**0.05** h0=**0**;

PAIRED YearsAtCompany \* YearsInCurrentRole;

**RUN**;

/\*As the p-value is less than 0.05, we reject the null hypothesis and conclude

that the mean difference in YearsAtCompany and YearsInCurrentRole is not 0\*/

/\*5\*/

/\*H0: Means of all department are the same.\*/

**PROC** **ANOVA** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable;

class Department;

model JobSatisfaction = Department;

means Department / hovtest welch;

**RUN**;

/\* Conclusion: Since the P-Value is greater than 0.05, we will accept the H0, meaning that the mean of all departments are the same. \*/

/\*6\*/

/\*H0: TotalWorkingYears is independent of MaritalStatus\*/

**PROC** **ANOVA** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable;

class MaritalStatus;

model TotalWorkingYears = MaritalStatus;

**RUN**;

/\* Conclusion: Since the P-Value is less than 0.05, we will reject the H0, meaning TotalWorkingYears is dependent of MaritalStatus \*/

/\*7\*/

/\*H0: Education is independent of JobRole\*/

**PROC** **ANOVA** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable;

class JobRole;

model Education = JobRole;

**RUN**;

/\* Conclusion: Since the P-Value is less than 0.05, we will reject the H0, meaning Education is dependent of JobRole \*/

/\*8\*/

/\*H0: The means of all BusinessTravel are the same. \*/

**PROC** **ANOVA** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable;

class BusinessTravel;

model JobInvolvement = BusinessTravel;

means BusinessTravel / hovtest welch;

**RUN**;

/\* Conclusion: Since the P-Value is greater than 0.05, we will accept the H0, meaning that the means of all BusinessTravel are the same. \*/

/\*Data Visualisation\*/

TITLE "Attrition vs Gender";

**PROC** **SGPLOT** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable PCTLEVEL=GROUP;

VBAR Gender / Group=Attrition1 STAT=PERCENT;

LABEL EMBARKED = "Employee Embarking Port";

**RUN**;

/\*Here we can see that there is very slight difference between male and female employees\*/

TITLE "Attrition vs JobSatisfaction";

**PROC** **SGPLOT** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable PCTLEVEL=GROUP;

VBAR JobSatisfaction / Group=Attrition1 STAT=PERCENT;

LABEL EMBARKED = "Employee Embarking Port";

**RUN**;

/\*Hence, as Job satisfaction increases, Employee Attrition = 0 which means that it is more unlikely for employees to leave the company\*/

TITLE "Attrition vs JobInvolvement";

**PROC** **SGPLOT** DATA=WORK.FINAL\_PROJ\_NT\_TaregtVariable PCTLEVEL=GROUP;

VBAR JobInvolvement / Group=Attrition1 STAT=PERCENT;

LABEL EMBARKED = "Employee Embarking Port";

**RUN**;

/\*Surprisingly, Employee Attrition is more when Job Involvement is high\*/

/\*Renaming few columns\*/

**DATA** WORK.FINAL\_PROJ\_NT\_ID;

SET WORK.FINAL\_PROJ\_NT\_TaregtVariable (RENAME= LogMonthlyIncome = MonthlyIncome Rename=LogJobHop = JobHop);

**RUN**;

/\*Check column names to enter in next step\*/

**PROC** **CONTENTS** DATA=WORK.FINAL\_PROJ\_NT\_ID OUT=WORK.FINAL\_PROJ\_NT\_ID2 (KEEP=NAME);

**RUN**;

/\*Building the model\*/

**PROC** **LOGISTIC** DATA=WORK.FINAL\_PROJ\_NT\_ID;

CLASS Attrition1 (PARAM=REF) BusinessTravel (PARAM=REF) Department (PARAM=REF)

EducationField (PARAM=REF) Gender (PARAM=REF) JobRole (PARAM=REF)

MaritalStatus (PARAM=REF) OverTime (PARAM=REF);

MODEL Attrition1 (Event = '1') = Age BusinessTravel DailyRate Department

DistanceFromHome Education EducationField EnvironmentSatisfaction

Gender HourlyRate JobHop JobInvolvement JobRole JobSatisfaction

MaritalStatus MonthlyRate MonthlyIncome NumCompaniesWorked

OverTime PercentSalaryHike PerformanceRating RelationshipSatisfaction

StockOptionLevel TotalWorkingYears TrainingTimesLastYear

WorkLifeBalance YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion

YearsWithCurrManager / SELECTION=Backward SLS=**0.05**

LINK=LOGIT

CLODDS=PL

ALPHA=**0.05**;

OUTPUT OUT = WORK.FINAL P=NEW;

**RUN**;

/\* Looking at the odds ratio and the likelihood estimates we can conclude that:

- Using backward selection of variables resulted in excluding 10 variables

based on the Pr>ChiSq date.

- We used reference coding for the categprical variables.

- Based on the above data, below are examples of some results/conclusions.

1. Female employees are (1-0.677) = 0.33 times that is 33% less likely to leave

the company(attrition).

2. With 1 increase in Age, 4% employees are less likely to leave

the company(attrition).

3. Sales Representative is more likely to leave the company than any other Job Role employee.

4. Data for YearsAtCompany and YearsSinceLastPromotion Suggests that with 1 increase in

in YearsAtCompany and YearsSinceLastPromotion, 100% employees are likely to leave the company.

/\*