**Capstone Project 1 - Milestone Report**

**1. Problem Definition**

The employee attrition and turnovers are always issued for either small or large businesses. The biggest challenge in this case is retention of valuable employees in order to provide the continuity of steady state organizations. The department of the human resources is responsible for not only to set reasonable immediate pay for employees but also to seek motivational factors to raise employees performance, which has a great contribution on business. By identifying factors that have impact on employees’ attrition and performance, the company will set HR policies during the decisions making process. Otherwise, the company would spend more time and money to hire a new employee.

**2. Identify your client**

The department of human resources of IBM (International Business Machine Corp.), Technology Company, represents the client of this project. Their concern is how they can detect the top key drivers of attrition and their importance by size to predict the attrition in order to get some precautions before an employee is more likely to leave from the company.

**3. Describe your data set, and how you cleaned/wrangled it**

The data, created by IBM data scientist, related to the IBM Human Resources (HR) acquired from kaggle.com, but the main data is still available in https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/. Only one data set is available for this project. Before exploring the data set, some initial examinations were applied during the data wrangling process. Data consists of 35 columns and 1470 observation. Numerical, categorical and binary variables were identified. Some columns such as “Over18”, “EmployeeCount” and “StandardHours” contain one type of value, so they were deleted. There were no missing values or non-unique values examined. Some categorical variables are relabeled by using explanation regarding to coded numbers.

**4. Explain your initial findings**

The data was split into two parts based on attrition or no attrition. Different visualization technics such as scatter plots, histograms and bar plots were applied on individual or group of variables. Before applying those technics, data was filtered by a dummy variable named 'Attrition'. Visualization technics were built on all data related to employees who left the company. The features of employees in attrition were learnt by count plots, bar plots scatter plots. Also, plot of correlation map were distracted for numerical variables.

Based on pearson-r correlation results:

# Age highly correlated with TotalWorkingYears(=0.7)

# JobLevel highly correlated with TotalWorkingYears(=0.7)

# JobLevel highly correlated with MonthlyIncome(=1.0)

# MonthlyIncome highly correlated with TotalWorkingYears(=0.8)

# PercentSalaryHike highly correlated with PerformanceRating(=0.8)

# YearsAtCompany highly correlated with YearsInCurrentRole and YearsWithCurrManager (=0.8)

# YearsAtCompany highly correlated with YearsSinceLastPromotion(=0.8)

# YearsCurrentRole highly correlated with YearsWithCurrManager(=0.7)

**According to findings extracted from histograms and bar plots:**

* Employees who had overtime from Sales department having around 6000 monthly income are more likely to leave from other departments.

# Employees who had JobLevel 5 from Research & Development Department having around 20000 monthly income are more likely to leave employees from Sales department having similar categories.

# Low job involvements from Sales and Research & Development department are more likely to leave.

# Males having around 5500 monthly incomes from Sales department are more likely to leave from females having 4500 monthly incomes in the same department.

# Job Level 1 in Human Resources having 15 miles distances from home are more likely to leave. Distance From Home in high job levels can be an issue for people from Sales Department.

* Average Monthly Rate for females is %4 higher than Males. We know that attrition in males is higher than females’, so monthly rate can be an influential factor for that reason.

**Based on the box plots of different categories('Gender', 'MaritalStatus', 'JobRole') towards 'DistanceFromHome':**

* 'Managers' who are male, married and around 15 miles far from home are likely to leave.
* 'Manufacturing Director' who are male, married and above 20 miles far from home are more likely to leave
* 'Healthcare Representatives' who are male, married and above 20 miles far from home are more likely to leave.
* 'Sales Executives' who are female, married and around 18 miles far from home are more likely to leave than males and single females.

Note: People who are 'Manager', 'Manufacturing Manager', 'Healthcare Representatives' and 'Sales Executives' job roles have high monthly income (above $7000).

**According to hypothesis testings of two sample t-tests, there are some significant relationships between some variables:**

# Is Age a significant factor for an employee of stay or left?

# Is distance from home a significant factor for an employee of stay or left?

# Is Total Working Years a significant factor for an employee who was in attrition or not?

# Is Total Years Since Last Promotion a significant factor for an employee who was in attrition or not?

Based on t-test statistics and p-values regarding questions mentioned above, there are significant relationships between age & attrition, distancefromhome & attrition, TotalWorkingYears & attrition. However, based on the p value, which was greater than .05 , there is no significant relationship between YearsSinceLastPromotion & Attrition.