MGT 6203 Final Report: Analysis of IBM Employment Factors that Influence Attrition

Team #: 38

Team Members:

1. Jack Lusk; EdX username: (Jlusk | jaklusk@gmail.com)

Jack studied Industrial Engineering at the University of Central Florida and has worked in different positions in Industrial Engineering since 2015. He is currently working as a Senior Simulation Consultant creating discrete event simulations for clients.

2. Sasakorn Phanitsombat; EdX username (KaoSasakorn | sasakorn.kao@gmail.com)

Sasakorn graduated with a Bachelor of Business Administration, majoring in finance. She is currently a program manager in a consumer technology company with a focus on e-commerce operations.

Introduction

Project Overview

This project aims to leverage IBM's employee data to build a predictive model that can help identify the factors influencing employee longevity and provide insight for improving employee retention strategies. For example, if job satisfaction is found to be a driver for attrition, companies can focus on improving work-life balance, career development opportunities, or employee recognition programs. In addition, hiring managers can refine their hiring processes based on attrition indicators.

Problem Statement

Employee attrition is a significant challenge faced by organizations today. High turnover rates not only impact productivity and team dynamics, but also result in substantial costs associated with hiring, training, and onboarding new employees.

Data Overview

IBM HR Analytics Employee Attrition & Performance | Kaggle https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

The project will utilize a dataset with 1470 rows of employee data from IBM and 32 indicator variables.

Age	Attritio	Busines	DailyRa	Educati	Educati	Environ	Gender	HourlyF	JobInvo	JobLeve	JobRole	JobSatis	Marital	Monthl	Monthl	NumCo	Over18	OverTir	Percent	Perform	Relatio	Standar
41	Yes	Travel_Rar	1102	2	Life Scienc	2	Female	94	3	2	Sales Exec	4	Single	5993	19479	8	Y	Yes	11	3	1	80
49	No	Travel_Fre	279	1	Life Scienc	3	Male	61	2	2	Research S	2	Married	5130	24907	1	Y	No	23	4	4	80
37	Yes	Travel_Rar	1373	2	Other	4	Male	92	2	1	Laboratory	3	Single	2090	2396	6	Y	Yes	15	3	2	80
33	No	Travel_Fre	1392	4	Life Scienc	4	Female	56	3	1	Research S	3	Married	2909	23159	1	Y	Yes	11	3	3	80
27	No	Travel_Rar	591	1	Medical	1	Male	40	3	1	Laboratory	2	Married	3468	16632	9	Y	No	12	3	4	80
32	No	Travel_Fre	1005	2	Life Scienc	4	Male	79	3	1	Laboratory	4	Single	3068	11864	0	Y	No	13	3	3	80
59	No	Travel_Rar	1324	3	Medical	3	Female	81	4	1	Laboratory	1	Married	2670	9964	4	Y	Yes	20	4	1	80
30	No	Travel_Rar	1358	1	Life Scienc	4	Male	67	3	1	Laboratory	3	Divorced	2693	13335	1	Y	No	22	4	2	80

The dataset includes the following indicator variables:

1. Job satisfaction: Job satisfaction rating out of 5
2. Years at the company: Number of years the employee has been with the company
3. Performance rating: Performance rating out of 5
4. Salary: Monthly salary
5. Percent raise: Salary raise from last year's salary
6. Several demographic information such as age, gender, and education level

The independent variable is attrition.

Research Questions

What are the influential indicator variables that contribute to employee attrition based on IBM's employee data?

1. What is the relationship between job satisfaction and employee attrition?
2. How does the number of years an employee has been with the company impact their likelihood of attrition?
3. Is there a correlation between performance ratings and employee attrition?
4. Does salary or play a significant role in employee attrition?
5. Does overtime have a significant impact in employee attrition?

Anticipated Conclusion

Certain indicator variables will be much more effective for attrition rates than others. Based on a cursory review of the data, we suspect that job satisfaction, years at the company, salary, and percent salary raise will have a substantial impact on whether an employee stays or leaves. Also, there will be certain indicators that are more influential and cheaper to improve. Companies should focus on these "low-hanging fruit" as they have the highest impact for the lowest cost.

Literature Review

1. S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry", *Imperial Journal of Interdisciplinary Research*, vol. 2, no. 8, 2016.

Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry | Semantic Scholar

The authors focus on the elements of positive work environment, developing, and retaining talent in retail environment. It is helpful to get another industry other than IBM.

2. G. K. P. V. Vijaya Saradhi, "Employee churn prediction", *Expert Systems with Applications*, vol. 38, no. 3, pp. 1999-2006, 2011.

Employee churn prediction - ScienceDirect
This paper looks at customer churn and uses elements for improving loss of customers for loss of employees. I find this an interesting approach that should be further investigated.

3. D. A. B. A. Alao, "Analyzing employee attrition using decision tree algorithms", *Computing Information Systems Development Informatics and Allied Research Journal*, no. 4, 2013.

ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS | D. | Computing, Information Systems, Development Informatics and Allied Research Journal (iiste.org)

The dataset is 109 observations for workers at a Nigeria institution. Waikato Environment for Knowledge Analysis (WEKA) and See5 were used to generate a decision tree model. They mainly used demographics information.

Data Collection and Preprocessing

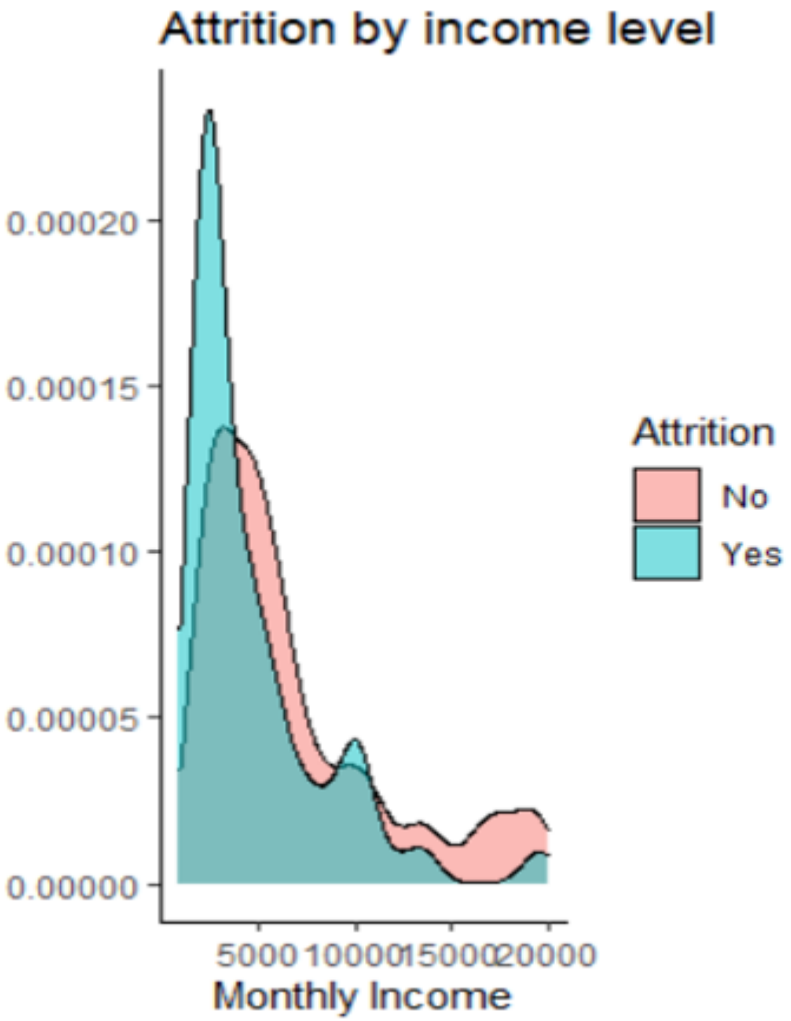
IBM HR Analytics Employee Attrition & Performance | Kaggle https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

The project will utilize a dataset with 1470 rows of employee data from IBM. The team performed a cursory data validation for blanks or invalid data and found the data has been processed. The variables BusinessTravel, Department, education, and EducationalField, OverTime, and Attrition were changed to categorical variables. Also, after looking at the four salary variables: DailyRate, HourlyRate, MonthlyIncome, and MonthlyRate the team found discrepancies and we are using MonthlyIncome for the models. The variables: EmployeeNumber, Over18, StandardHours were removed because they are insignificant.

Exploratory Data Analysis

In the exploratory data analysis (EDA), the main focus is to look at attrition rate with different indicator variable categories. We use a combination of graphs to visualize the results.

Attrition by Income Level

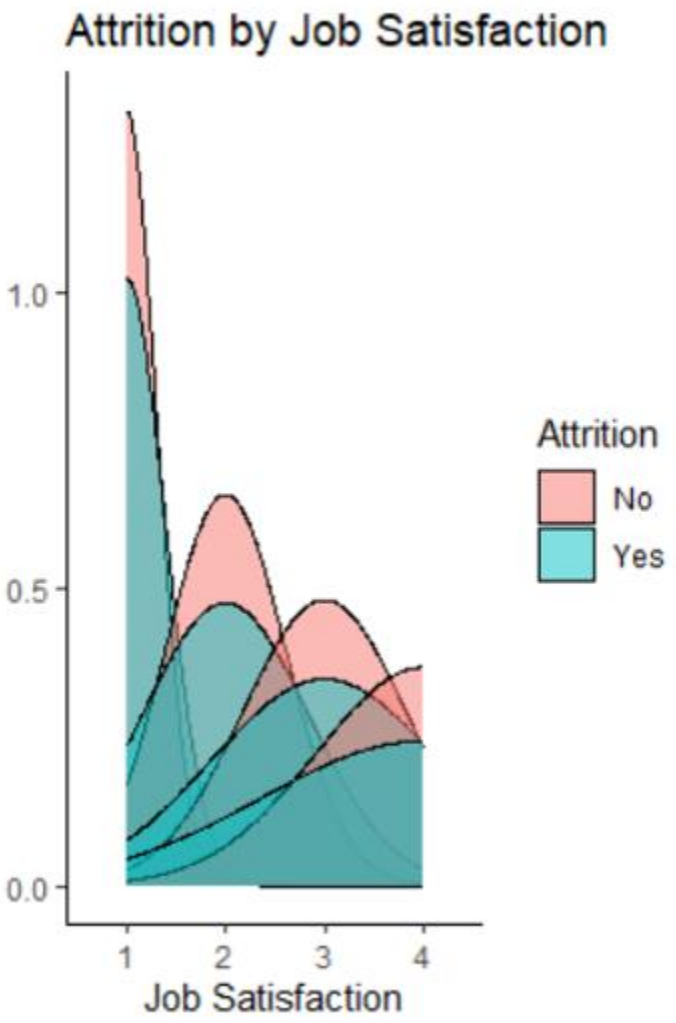


As we can see, attrition is much higher at lower monthly income. The levels are the highest where monthly income is less than $5000.

Attrition by Department

Attrition by Years at Company

Attrition by Job Satisfaction



Attrition by Overtime

Out of the employees that stayed 23.4% worked overtime and out of the employees that left 53.6% employees worked overtime. There seems to be a high correlation between overtime and attrition.

Methodology and Approach

We used supervised machine learning algorithms to model our data. The two supervised ML models that we built and tested were Support Vector Machine and Decision Tree. We also built a logistic regression model. The dataset was split into training dataset and test dataset at a 80%/20% split. Let's first look at the Support Vector Machine.

Support Vector Machine

We are using the radial kernal and the C-classification type.

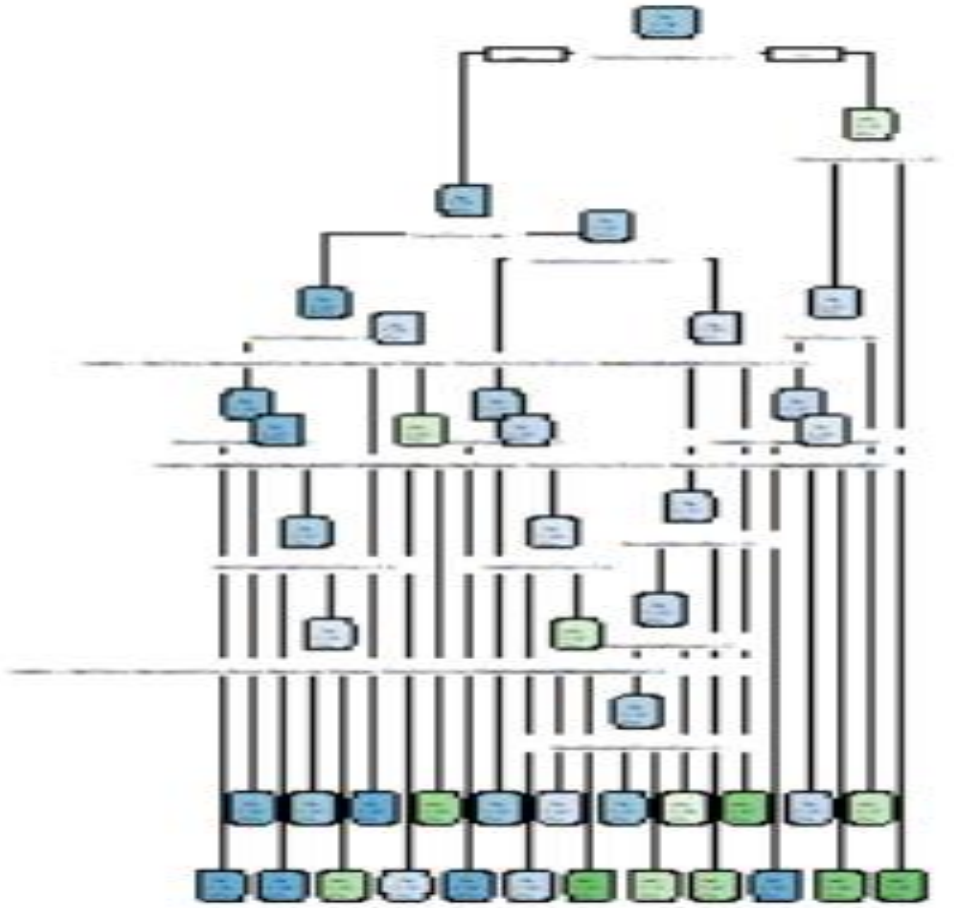
```
svm(formula = train_data$Attrition ~ ., data = train_data,
    type = "C-classification", kernel = "radial")
```

We trained the model with the training dataset. We ran predict with the test dataset and ran a confusion matrix with the caret package.

We found the accuracy to be 84.0%.

Decision Tree

We ran the decision tree model with the rpart package and training dataset. The main findings with the decision tree that it has main splits at TotalWorkingYears >= 1.5, OverTime= "No" , WorkLifeBalance >=2, JobSatisfaction >=2. This means those factors are significant and separate the dataset well.



We tested the model with the test dataset. The model accuracy was found to be 78.9%.

Logit Regression Model

We first change the response variable, Attrition, to 0s and 1s. We then use glm function and the main response variables in the model.

```
glm_model <- glm(Attrition~ Age + YearsInCurrentRole + OverTime +
DistanceFromHome+Education+EnvironmentSatisfaction+ HourlyRate+JobInvolvement
```

+ JobLevel+JobSatisfaction + MonthlyIncome + NumCompaniesWorked + PercentSalaryHike+ PerformanceRating +

TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +YearsAtCompany + YearsSinceLastPromotion + YearsWithCurrManager, data = train_data, family = "binomial")

The AIC is 813.47 (lower is better model performance). We also found the variables that were the most significant: OverTime, DistanceFromHome, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, NumCompaniesWorked, WorkLifeBalance.

More importantly, PercentSalaryHike, TrainingTimesLastYear, YearsAtCompany, JobLevel, HourlyRate, Education are not statistically significant.

Recommendations

Based on the exploratory data analysis, we found overtime is a large factor for attrition. Hence, from business standpoint, we should aim to optimize employees' working productivity and efficiency and minimize no. of redundant projects and hence number of hours employees working overtime. Productivity & Efficiency during the official working hours shall be our key focus. Human resource projects that will help with this focus can be for example, limiting no. of meetings which will allow employees to spend more time in thinking of high-impact work and meetings are only required when necessary.

Conclusion

We have deployed multiple models and compare their performances. We also transformed the data required to better suit the model methodologies we're applying. The most promising models that give us the highest accuracy suggest that overtime is the largest factor for attribution. From this on, we advise the company to initiate projects around increasing productivity & efficiency of employees during working hours to increase their attrition.

Future Work

With the rapid changes in industries and disruptions we see, it's unavoidable that company will need to put more hours per employee given limited resources. In the future work, we can explore the project around, how to increase or maintain employees' attrition given that they need to work over time – what are the compensation forms that would most contribute to employees' attrition.