Predicting Employee Attrition at IBM
By Kevin Ba Ross

Part 1 - Background

Losing highly talented employees can be catastrophic for a company. Many companies spend millions of dollars hiring and training employees, and these efforts go to waste when employees quit. Retaining an employee is one of the most important processes, so it is an utmost important goal for a business to know why employees are leaving. Is it because their pay is not competitive? Is it because of the working environment? Or is it because we made a wrong decision by hiring candidates from an irrelevant educational field in the first place?

This report will explore the factors that contribute to employee attrition at IBM. The dataset used in the analysis is the real data provided by three departments of IBM: sales, human resources, and research and development. The findings will be valuable for management at IBM to effectively allocate its resources to reduce its employee turnover and retain those highly talented workers.

This paper will begin by discussing the nature of the data in the dataset and the methodologies used to analyze the data. The paper will then analyze and evaluate the findings and the significance of each variable. After that, the limitations of the model recommendations on how the model can be improved will be discussed. Finally, the report will recommend managerial implications for the management of IBM.

Part 2 - Methodologies

This paper will analyze the data on 1,470 employees collected by IBM in 2019 by using mostly Python and Jupyter Notebook. To analyze the data, supervised machine learning will be used. Machine learning is the process of using computer algorithms and statistics to learn from the data, and then use that information to make judgments or predictions about the future data. The main difference between unsupervised and supervised machine learning is that, in supervised learning, the algorithm tries to understand the unlabeled dataset by extracting common characteristics and patterns from the data. However, with supervised machine learning, the data scientists feed the algorithm with outcomes (or the dataset that is labeled with answers) for the algorithm to learn from. These answers will be used as a key to evaluate the accuracy of the model.

In the dataset provided, the outcome is binary, as it is explicitly given whether an employee attrites or not, depending on certain characteristics of an employee, such as demographics, performance rating, salary, etc. Because the outcome (`Attrition_binary`) is given and because we are trying to predict what factors are significant in causing employees to attrite, the supervised machine learning will be more appropriate than the unsupervised one.

The logistic regression will be used to train and test the data; then the classification report will be created to assess the overall performance of the model. The logistic regression is suitable for this dataset as it has a binary outcome (such as 0 and 1). To analyze the significance of each variable, odds ratios will be used. With odds ratios, we can see how likely (or unlikely) the employee will attrite. This will provide a valuable piece of information for the management at IBM to see which areas they have to focus on to minimize employee attrition.

Part 3 - Findings

3.1) Findings on Overall Model Performance

By starting to explore the data of 1,470 IBM employees, about 16.1% of them attrite. This shows that the model can be imbalanced as the majority of the data are from those who do not attrite. As a result, the model may be better at predicting those who do not attrite than those who do. To prove this point, the confusion matrix and the classification report are run to assess this hypothesis. The table below shows the snapshot of the classification report.

	Precision	Recall	F1_Score
0 (Not_Attrite)	0.87	0.99	0.93
1 (Attrition)	0.75	0.17	0.27

As we can see from the classification report above, the precision score, recall score, and F1 score are higher in the `Not_Attrite` row. This finding indicates that the model did much better at predicting those who would not attrite than predicting those who will attrite.

Now, the report will explore even further by using the odds ratios to see how likely each employee will leave based on certain characteristics. By using the code from Professor Guilbeault, the odds ratios for all variables are shown below. Because of the large number of variables, this report will explain the odds ratios in relation to their theme in the following section.

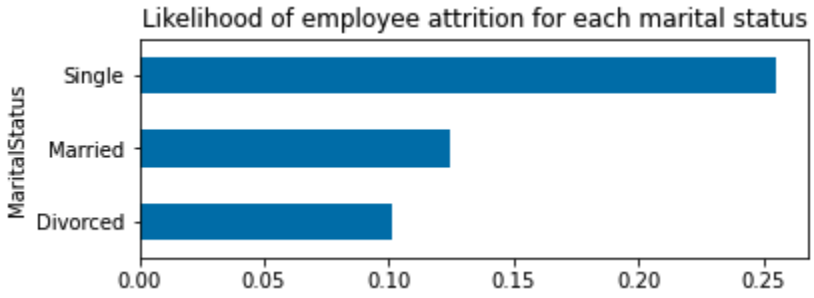
	Attrition_binary			
Predictors	*Odds Ratios*	*std. Error*	*CI*	*p*
Intercept	32.06	1.18	3.20 – 321.67	**0.003**
JobSatisfaction	0.72	0.07	0.63 – 0.83	**<0.001**
EnvironmentSatisfaction	0.72	0.07	0.62 – 0.83	**<0.001**
PercentSalaryHike	0.96	0.03	0.89 – 1.03	0.221
HourlyRate	1.00	0.00	0.99 – 1.01	0.817
Department=Research & Development	0.95	0.53	0.33 – 2.67	0.916
Department=Sales	1.56	0.55	0.53 – 4.57	0.414
Education	0.98	0.08	0.84 – 1.14	0.774
EducationField=Life Sciences	0.39	0.73	0.09 – 1.63	0.198
EducationField=Marketing	0.60	0.77	0.13 – 2.68	0.500
EducationField=Medical	0.35	0.73	0.08 – 1.45	0.146
EducationField=Other	0.44	0.79	0.09 – 2.08	0.302
EducationField=Technical Degree	0.74	0.75	0.17 – 3.21	0.692
Gender=Male	1.23	0.16	0.89 – 1.69	0.208
JobInvolvement	0.62	0.11	0.50 – 0.77	**<0.001**
MaritalStatus=Married	1.30	0.23	0.82 – 2.06	0.263
MaritalStatus=Single	3.58	0.26	2.16 – 5.93	**<0.001**
NumCompaniesWorked	1.13	0.03	1.06 – 1.21	**<0.001**
PerformanceRating	1.47	0.35	0.74 – 2.92	0.273
TotalWorkingYears	0.94	0.02	0.90 – 0.98	**0.008**
TrainingTimesLastYear	0.83	0.06	0.74 – 0.95	**0.005**
WorkLifeBalance	0.75	0.11	0.61 – 0.93	**0.009**
YearsAtCompany	1.09	0.04	1.01 – 1.16	**0.020**
YearsInCurrentRole	0.87	0.04	0.80 – 0.95	**0.001**
YearsSinceLastPromotion	1.17	0.04	1.09 – 1.26	**<0.001**
YearsWithCurrManager	0.88	0.04	0.81 – 0.95	**0.002**
MonthlyIncome	1.00	0.00	1.00 – 1.00	**0.003**
children	1.07	0.12	0.84 – 1.36	0.568
Manager_relation_satisfaction	0.85	0.07	0.74 – 0.98	**0.029**
Observations	1470			
R^2	0.267			

3.2) Findings on Demographic Information

Significant Variables More likely to attrite	**Significant variables** Less likely to attrite	**Insignificant variables** P-value > 0.05

`MaritalStatus=Single`	None	`Gender=Male`
		`MaritalStatus=Married`
		`Education`
		`EducationField (All)`
		`children`

Even though the bar graph below clearly shows that employees who are married and divorced are less likely to attrite than those who are single, the odds ratios show that only the variable `MaritalStatus=Single` is significant in predicting employee attrition at IBM.



Limitations and Recommendations

As single employees are more likely to attrite, it would be more helpful for IBM to collect data on 'Single but living together' or 'Does not want to declare.' If employees live with their life partners (not officially married), they may be more committed to staying in a particular city with their partner than those who are actually single. Another question that IBM should consider is has this data been collected for a long time but has not been updated? These recommendations could help IBM see a clearer picture on how marital status can affect the attrition rate.

3.3) Findings on Pay and Compensation

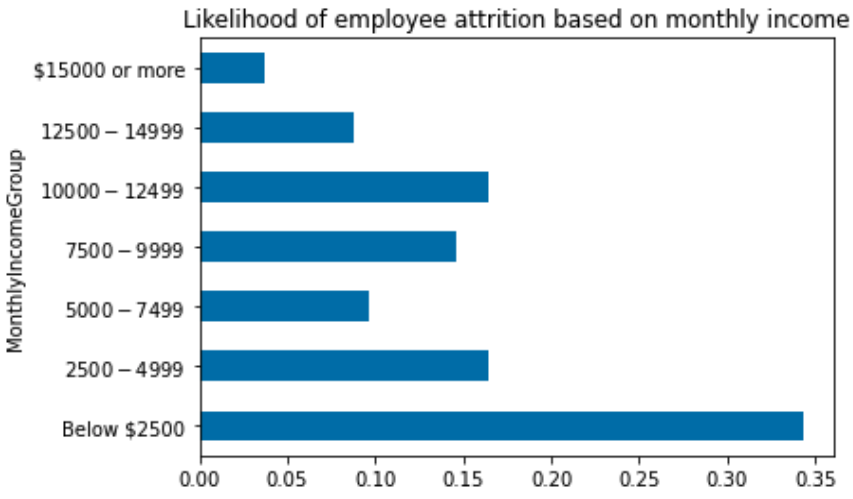
Significant Variables More likely to attrite	**Significant variables** Less likely to attrite	**Insignificant variables** P-value > 0.05
None	MonthlyIncome	PercentSalaryHike HourlyRate

The odds ratios show that the percent of salary increase and hourly rate are not significant in predicting attrition. Monthly income, however, is significant in predicting attrition as the p-value is less than 0.05. Because the odds ratios output in the table on page 3 is based on the monthly income at the unit of $1, I re-adjusted the data to make the unit of MonthlyIncome to become $1,000 to see a clearer effect on the likelihood of attrition. The following figure shows the summary of the re-run.

	Odds ratios	P > |z|
MonthlyIncome	0.920715	0.008

The odds ratio of 0.920715 implies that for every $1,000 that an employee earns in a month, the likelihood of their leaving the company is reduced by approximately 8% on average. However, when looking at the graph below, it gives a slightly different conclusion. It can be seen that the attrition rate is the highest for those with a monthly income of less than $2,500, which accounted for approximately 15% of the workforce at IBM during that time. Then, the attrition rate did not change much for more than three quarters of IBM employees, whose income group was between $2,500 and $14,999, according to the figures below.

Monthly income group	Number of employee (%)
$15,000 or more	133 (9%)
$12,500 - $14,999	57 (4%)
$10,000 - $12,499	91 (6%)
$7,500 - $9,999	130 (9%)
$5,000 - $7,499	310 (21%)
$2,500 - $4,900	525 (36%)
Below $2,500	224 (15%)
Total	1470 (100%)



Limitations and Recommendations

It is not surprising that employees with low income are more likely to leave the workplace to the one that they can earn higher pay. But can IBM make the decision right away to give the minimum monthly income at $2,500 just to reduce attrition? The answer to this question depends on several factors. IBM management has to ask whether such employees can get better pay elsewhere with the same level of skills that they currently have. If there are no other places that give higher salary for the same level of skills or if there are no other jobs that give the same level of satisfaction as the one at IBM, IBM may not have to increase the monthly income of an employee. Even though the dataset provides the data on departments, the dataset lacks information on the position that the employee holds and the industry's average pay of that position. If IBM collects these 2 pieces of data, it may be able to make a better decision on the pay scale of the employee.

Although this dataset has some limitations, there are several management implications that IBM can do to reduce employee attrition. This will be discussed in the 'managerial implications' section.

3.4) Findings on Employee Performance at Work

Significant Variables More likely to attrite	**Significant variables** Less likely to attrite	**Insignificant variables** P-value > 0.05
None	`JobInvolvement`	`PerformanceRating`

These two variables are assigned by the manager on a scale of 1 to 4. The higher score on the scale, the better the involvement and performance of the employee. The result shows that job involvement is a significant predictor of employee attrition, whereas the performance rating is not. The odds ratio of the variable `JobInvolvement` is 0.62, implying that if employees can increase their job involvement rating by 1 unit, the likelihood of attrition decreases by 38%.

Limitations and Recommendations

This finding can be useful for management at IBM to try to get employees to be more involved in their jobs. However, there are some limitations to this finding. First, we cannot be certain whether employees want to attrite, so they become less involved in the job. Or do the employees get discouraged when seeing their job involvement score and they want to attrite?

To clear up this complication, IBM should collect more data to see the factors that cause managers to rate employees in that way. The employee KPI has to be clear to see what determines involvement in the job. Is it the number of hours that they work? Is it the number of projects that they do? In addition, managers may miss out some employees who may be heavily involved in the project, but managers do not see it if the team size is very large, indicating proximity bias. It may also be helpful for IBM to have the rating from the colleagues in the team that the employee is working with, then average all the results to determine employee involvement to reduce bias.

Another recommendation is to perform a time series analysis to see whether employee involvement increases over time. Those with a rating of 3 (but 4 last period) may be more worrisome than those with a rating of 3 (but 2 last period). By analyzing the trend, IBM can get a clearer picture of employee performance and predict the likelihood of attrition accordingly.

3.5) Findings on Work Environment and Employee Satisfaction

Significant Variables More likely to attrite	**Significant variables** Less likely to attrite	**Insignificant variables** P-value > 0.05

None	`JobSatisfaction`
	`EnvironmentSatisfaction`
	`WorkLifeBalance`
	`Manager_relation_satisfaction`
		`Department=R&D`
		`Department=Sales`

The odds ratios show that factors relating to the work environment (job satisfaction, environment satisfaction, work-life balance, and manager relation satisfaction) are indeed significant predictors of employee turnover. Not surprisingly, the happier employees are more likely to stay with the company that they are happy to work with.

Although the odds ratios do not show whether the department that employees belong to is significant in predicting attrition, simply calculating the mean of 0 (no attrition) and 1 (attrition) from each department reveals in the graph below that those in the Human Resources Department and Sales Department are slightly more likely to attrite than those in Research and Development Department.

Limitations and Recommendations

It is important to note that variables related to satisfaction are self-reported by IBM employees on a scale of 1 (highly dissatisfied) to 4 (highly satisfied). There is the possibility of bias in this practice as the rating is very subjective. Two employees can have the same level of satisfaction but one could rate a 3 and the other a 4. This is very difficult to measure. In my opinion, since variables relating to satisfaction play a crucial role in predicting attrition, it would be worthwhile for IBM to spend effort in collecting data in a time-series manner to see the trend on how satisfied employees are with their job and their working environment.

It is also important to find out what factors lead to satisfaction in each of the categories above. Is it because of flexible workhour? Is it a safe working environment? Is it the psychological safety they have? It is difficult to tell from the given dataset alone. To understand what contributes to employee satisfaction, there are further steps that IBM can do to understand their employees better. This will be discussed in the 'managerial implications' section.

3.6) Findings on Career Progression

Significant Variables More likely to attrite	**Significant variables** Less likely to attrite	**Insignificant variables** P-value > 0.05

`NumCompaniesWorked`
`YearsAtCompany`
`YearsSinceLastPromotion`
	`TotalWorkingYears`
	`TrainingTimesLastYear`
	`YearsInCurrentRole`
	`YearsWithCurrManager`
		None

The results after running odds ratios suggest that if employees work for a long time at IBM but have not been promoted for a long period of time, they are more likely to attrite. However, if IBM provides them with training and a clear career path, they are more likely to stay. It is also important to note that even though employees are more likely to stay if they have been promoted, they are also more likely to stay if they are in the same role with the same manager.

This finding may seem contradictory, but one possible explanation is they may prefer to do the same thing, but on a larger scale, or they may want to be recognized with promotion. The section of 'managerial implications' at the end of the paper will explain additional steps IBM can take to understand the motivation of its employees.

Limitations and Recommendations

It is also helpful for IBM to collect data on the previous industry that the employees are working in to determine whether such employees should be in a more senior position, or they are career switchers. For example, an employee could work with 5 companies as a sales representative, but it is his first year in human resources. This will help IBM see a clearer picture on the amount of pay, or the additional training that this employee may need.

Part 4 - Managerial Implications

To summarize, the important themes emerged from the analysis above are monthly income (for those with income under $2,500), job involvement, employee satisfaction, and career progression opportunities. The combination of these findings yields several useful managerial implications.

Firstly, the finding indicates that those with income under $2,500 are more likely to attrite. If IBM does not want to lose such employees, why not train them to become more skillful at their job? The data above also indicates that career advancement opportunities and training are significant factors in reducing attrition. If IBM shows its employees, especially those on the lower-paid scale, that IBM has a clear career path for them and they can have good access to training (the training of which is hard to get elsewhere) that can increase their salary, such employees will be less likely to leave IBM.

Secondly, as employee satisfaction significantly reduces attrition, it is also important to find out what factors make them satisfied. To do so, IBM can allocate a group of representatives to help employees solve problems or to voice grievances they have in the form of anonymous online chat. Even if they don't come to online chat, IBM can have the survey for them to fill in every week regarding how they feel with their jobs; what additional support they need; and what type of training they need. This can, for example, be in the form of a pop-up on the screen every time the employees sign in to their work station, or surveys via email. Alternatively, IBM can encourage its employees to spend 5 minutes a day writing anonymous daily journals online on what they feel about the work each day.

By doing so, IBM can collect a lot of qualitative information and use unsupervised machine learning to cluster their results into groups to better understand the data. IBM can also analyze text and analyze sentiment to see co-occurrence themes or to see whether there are more positive than negative words, or if there are any trends that IBM should be aware of. This will enable IBM to see the trend in a time series manner and take action accordingly to prevent attrition.

However, there are many more factors to consider. The employees may earn a lot of money at IBM and are very happy with the work environment and their colleagues, but they may have to leave the company because their families move to another country, or they see a new business opportunity that they want to become an entrepreneur rather than having to work at IBM. That is why it is important for IBM to find out other causes beside those discussed above.