Copenhagen
Business School
HANDELSHØJSKOLEN

What are the factors of employee attrition?

A case study of IBM HR Data

Econometric Analysis of Firm Data (CCMVV2401U)

Written Exam

Student ID number: 158370

Examiner: Ralf Andreas Wilke

Submission Details

Date of submission: 27-02-2024

Number of characters: 26.828

Number of normal pages: 14

Abstract

This paper examines potential factors which are influencing retainment of employees in a cross sectional, econometric analysis. It uses Linear Probabilistic Model which is a binary extension of the Ordinary Least Squares and a Logistic Regression (Logit). It identifies that the magnitude and significance of various predictors of employee departure. The analysis reveals that age, distance from home, overtime, monthly income, travel frequency, department type, job role, and personal job satisfaction metrics (including environment satisfaction, job involvement, and work-life balance) significantly affect attrition rates. The contribution of this research is the focus on the identification of factors that lead to attrition rather than focusing on the prediction of it.

1. Introduction

As the modern labour markets are dynamic and more open internationally it is crucial for companies to be aware of the factors that play a role in retaining its employees. As whether an employee will stay with the company is an aspect that will directly impact the organisational efficiency and the continuity of the company. Now with the wide array of employment possibilities, especially with post-pandemic growing remote work culture (Kane et al., 2021), the impact of attrition poses a challenge for human resources department that ought to look into the factors that are contributing to retainment of employees as lack of it might have an impact on the continuity as well as stability of the company. Moreover, understanding the factors impacting the will to stay of employees can benefit the HR department to foster an environment where the employees would like to stay and offer the developments as well as amenities that they are currently lacking.

This research aims to determine factors and the magnitude of their impact on the employee retention based on synthetic data (IBM HR Analytics Employee Attrition & Performance, 2017) created by IMB Data Scientists. The research will focus on econometric analysis employing a special case of Ordinary Least Squares (OLS) which is designed for a binary variables - Linear Probabilisitic Model (LPM) and a Logistic Regression (Logit)

The contribution of this study lies in the existing knowledge on employee attrition providing empirical insights into which factors have the biggest impact on leaving the company. Moreover, the research space is dominated with prediction of employee attrition

rather than identifying factors causing the attrition itself. Furthermore, the specific architecture of data itself plays a crucial rule in its contribution, as there is an array of qualitative and quantitative variables, where the qualitative are expressed as ordinal nominal variables. The nuanced nature of employment data, characterised by ordinal variables that represent hierarchical or ordered categories without a fixed scale, presents analytical challenges that linear models may not fully address.

2. Research Question

This paper will answer a question *what are the most influential factors in a discussion of employee attrition?*

3. Literature Review

Das and Baruah (2013) provide a comprehensive literature review of the factors influencing employee retention. They offer a broad analysis that collects various academic papers which allow for a better understanding of the dynamics of the workforce. In their analysis they find that Gardner et al. (2004) discover high influence of the compensation on whether an employee will stay within the company. However, it is also mentioned that higher pay is not the universal solution as the strategic implications of using a competitive compensation package is crucial as it has impact on the organisational framework as well as the loyalty of employees. Furthermore, Das and Baruah (2013) draw attention to the significance of recognition in fostering a loyal workforce, as originally noted by Agarwal (1998). Recognition is not only characterised by promotions or other material aspects typically associated with positive job feedback but also the non-material acknowledgments which further affirm the internal value of the worker, what further translates to commitment to the organisation. Career advancement opportunities also signify a critical aspect of the retention of employees which positively correlate with the satisfaction (Pergamit & Veum, 1999). In recent years, Work-Life Balance has increasingly become more important to workers. While already in 2004 Hyman and Summers found that the employees that were asked to work outside of their working hours have been subjected to more stress and exhaustion, which could be a cause for attrition of employees. The work environment plays a substantial role in employee satisfaction and retention. Zhang (2016) mentions that, among

others, workload, type of company culture and flexible working hours are one of the most important *personal* factors in the employee turnover.

4. Data

4.1 Preliminary analysis

The data used for this analysis was found on Kaggle, which is also the official source listed on the IBM website[1]. The dataset is synthetic and it concerns the employee data. It was developed by a team of data scientists at IBM for learning purposes of different Machine Learning models.

The preliminary analysis of the dataset reveals that there are 1470 observations across 35 different variables. This preliminary analysis performs robustness checks for missing data and does not identify any. The variables in the dataset are numeric, categorical and ordinal-nominal. They present a broad spectrum of employees characteristics as well as their intrinsic perceptions of their working conditions. It covers, among others, the basic demographic information such as gender, age or level of achieved education, while the before mentioned

Summary Statistics

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Age	18.00	30.00	36.00	36.92	43.00	60.00
DailyRate	102.0	465.0	802.0	802.5	1157.0	1499.0
DistanceFromHome	1.000	2.000	7.000	9.193	14.000	29.000
Education	1.000	2.000	3.000	2.913	4.000	5.000
EmployeeCount	1	1	1	1	1	1
EmployeeNumber	1.0	491.2	1020.5	1024.9	1555.8	2068.0
EnvironmentSatisfaction	1.000	2.000	3.000	2.722	4.000	4.000
HourlyRate	30.00	48.00	66.00	65.89	83.75	100.00
JobInvolvement	1.00	2.00	3.00	2.73	3.00	4.00
JobLevel	1.000	1.000	2.000	2.064	3.000	5.000
JobSatisfaction	1.000	2.000	3.000	2.729	4.000	4.000
MonthlyIncome	1009	2911	4919	6503	8379	19999
MonthlyRate	2094	8047	14236	14313	20462	26999
NumCompaniesWorked	0.000	1.000	2.000	2.693	4.000	9.000
PercentSalaryHike	11.00	12.00	14.00	15.21	18.00	25.00
PerformanceRating	3.000	3.000	3.000	3.154	3.000	4.000
RelationshipSatisfaction	1.000	2.000	3.000	2.712	4.000	4.000
StandardHours	80	80	80	80	80	80
StockOptionLevel	0.0000	0.0000	1.0000	0.7939	1.0000	3.0000
TotalWorkingYears	0.00	6.00	10.00	11.28	15.00	40.00
TrainingTimesLastYear	0.000	2.000	3.000	2.799	3.000	6.000
WorkLifeBalance	1.000	2.000	3.000	2.761	3.000	4.000
YearsAtCompany	0.000	3.000	5.000	7.008	9.000	40.000
YearsInCurrentRole	0.000	2.000	3.000	4.229	7.000	18.000
YearsSinceLastPromotion	0.000	0.000	1.000	2.188	3.000	15.000
YearsWithCurrManager	0.000	2.000	3.000	4.123	7.000	17.000

Table 1

[1] https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/?mhsrc=ibmsearch_a&mhq=IBM HR Analytics Employee Attrition & Performance

intrinsic perceptions cover aspects like job satisfaction, work-life balance or environment satisfaction.

In Table 1 can be seen a summary of the nominal and ordinal nominal variables in the dataset. It can derived that the mean age of the employees is 36.92 years old. Moreover, the range of monthly income is quite vast, from $1009 to $19999, which is highlighting the possible economic disparities within the workforce. The data also presents that employees range from IBM being their first job, while the maximum for the companies worked before is nine, with a mean of 2,7 which could show that the employee pool, given the age mean, is at the beginning of their career. Moreover, looking at the Years At Company variable, where the mean is 7 years, it is above average, across all regions according to Choughari (2022).

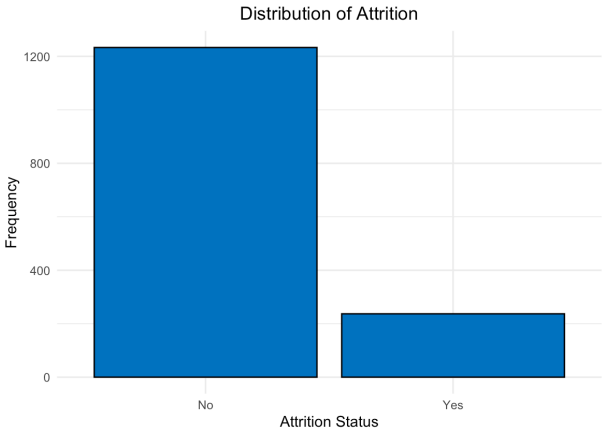


Figure 1: Distribution of Attrition

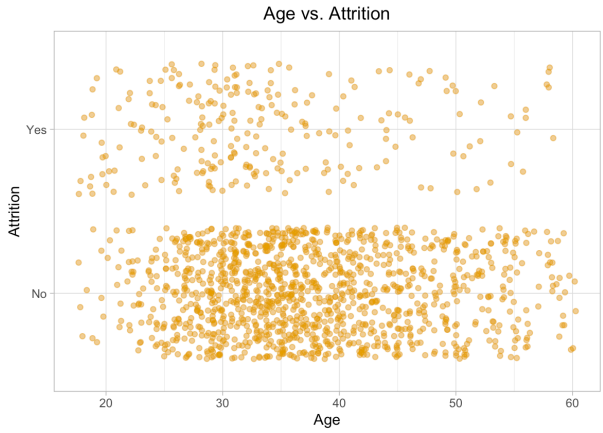


Figure 2: Scatter Plot Attrition vs. Age

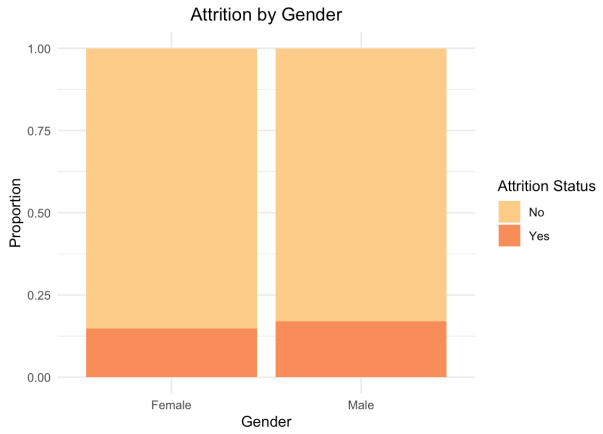


Figure 3: Attrition vs. Gender

When it comes to the intrinsic variables, Environment Satisfaction, Job Involvement, Job Satisfaction, Performance Rating, Relationship Satisfaction and Work-Life Balance is becomes clear that the mean of them is ‘above-average’ as IBM HR Analytics Employee Attrition & Performance (2017) identifies the levels a follows: 1 ‘Low’, 2 ‘Medium’, 3 ‘High’, 4 ‘Very High’. Each of these variables’ mean being at the 2.7 level shows that employees are overall happy with their employment. Lastly, an important variable to consider is Education, which as previously, is an ordinal ordered variable where 1 signifies 'Below College’, 2 - ‘College’, 3 - ‘Bachelor’, 4

- ‘Master’ and 5 - ‘Doctor’ respectively. The median of this variable is 3, what shows that most of the employees do hold a bachelor degree, while the mean is 2.9 which presents that the eduction is ever so slightly skewed towards less educated employees.

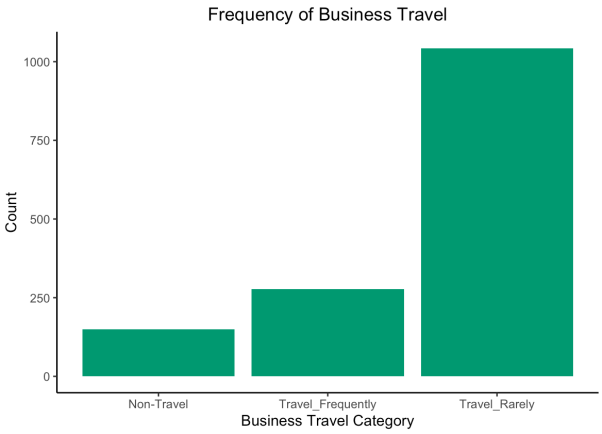


Figure 4: Bar Chart Business Travel

4.2 Exploratory Data Analysis

The EDA focuses on finding relationships between different variables and Attrition, which will be the variable of interest in the regression sections of this paper. On Figure 1 can see that the inherit distribution of Attrition presenting that there are around 200 employees which are leaving the company, while the majority decides to stay. In Figure 2 can be seen that there might be a loose cluster forming of the employees who left and are in their mid-20s until mid-30s. Furthermore, when looking into Attrition by Gender (Figure 3) it can be concluded that slightly more males prefer to leave the company compared to females. However, the number seem to be marginal. In Figure 4 can be seen that most of the employees rarely travel, with around 8% does not travel at all. Moreover, in the correlation chart (Figure 5) (larger version can be seen in the Appendix) it can be seen that the Total Working Years and Job Level are strongly and positively correlated what is an intuitive outcome as the longer one works the higher position they acquire within the company. Next, the previously mentioned Total Working Years is also positively correlated with the Monthly Income implying that employees with more years of total work experience tend to have higher monthly incomes which is as well an intuitive conclusion. Moreover, a positive correlation between the Percent Salary Hike and the higher Performance Rating suggests that there is a reward system in place for employees which are performing well. And

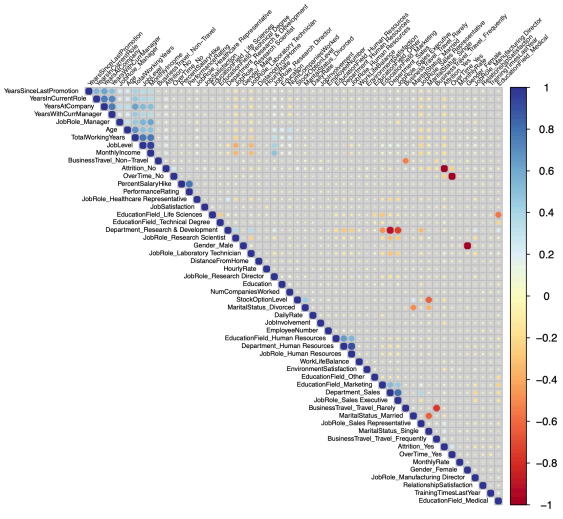


Figure 5: Correlation plot of all variables (dummy version)

lastly, there is a negative correlation between Job Role as a Sales Representative and Total Working Years which could indicate that the sales representative role is often occupied by individuals with fewer total working years, possibly suggesting it's a common entry-level position or a role with higher turnover.

5. Methodology

The econometric approach adopted in this study involves two main methodologies: Ordinary Least Squares (OLS) - Linear Probabilistic Model and a Logistic Regression. In the data there are multiple columns which are ordinal but nominal. Based on Torra et al. (2006) one of the approaches to deal with variables which are nominal in nature however, the distance between the interval between the bordering values is not known it is to ignore the unknown internal and treat the variable as nominal. As the alternatives suggested to be used instead, go beyond the scope of this paper, this approach will be used. Moreover, with the usage of the dummy variables, not all generated dummies were used, firstly to avoid 'so called dummy variable trap' (Wooldridge, 2019) but also from the simple space constraints. The dataset with dummy variables had 57 variables so not all will be used.

5.1 OLS - Linear Probabilistic Model

The Linear Probability Model (LPM) is a regression used to estimate a probability of a binary outcome based on one or multiple independent variables. It is considered a special case of an Ordinary Least Squares (OLS) in which the dependent variable is a dummy, taking values of either one or zero. Given that the LPM's task is to estimate the probability of the outcome being one based on the set of independent variables. Therefore, the equation of this model is as follows:

$$P(Y = 1 \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where Y represents the dependent variable, $X_1, X2,...,X_n$ the independent variables. Moreover, α represents the intercept while the $\beta_1, \beta_2, \ldots, \beta_n$ represent the coefficients which measure the influence of the independent variable on the dependent one. As it is a model based on the OLS it is crucial to discuss its approach. The OLS method seeks to minimise the sum of the squared residuals - differences between the observed dependent variables and the

predictions of the model (Wooldridge, 2019). This technique is one of the most widely used ones in econometrics.

There are five assumptions of Multiple Linear Regression (Wooldridge, 2019):

1. Linear Parameter: The model is linear in parameters, what implies that the relationship between the independent and dependent variables is linear with respect to the coefficients.
2. Random Sampling: The data consist of random sample of n observations, which ensures that the population is well represented and the estimates are unbiased.
3. No Perfect Multicollinearity: None of the independent variables are perfectly correlates with themselves. If that is the case the it is impossible to distinguish which of the variables affects the dependent variable which causes infinite or undefined estimates of coefficients.
4. Zero Conditional Mean: The expected value of the error term conditional on the independent variables is zero. This assumption means that the errors are uncorrelated with the independent variables, ensuring that the OLS estimates are unbiased.
5. Homoskedasticity: The error u has the same variance given any value of the explanatory variables: $Var(u \mid x_1, \ldots, x_k) = \sigma^2$

In the case of Linear Probabilistic Model the Homoskedasticity assumption is violated by default, but there are techniques to counteract that violation. Among others it is usage of a White estimator, which corrects the estimations (Linear Probability Model, n.d.). As this is a linear model of probability, and the linear predictions are not bound between 0 and 1, one needs to be careful when assessing the models (Linear Probability Model, n.d.).

5.2 Logistic Regression - Logit

The Logistic Regression is a model specifically designed for binary outcome variables. It models the log odds of the probability of the outcome as a linear function of the predictor variables, addressing some limitations of the LPM by ensuring that predicted probabilities always fall between 0 and 1. To ensure that interval the logistic function is used which has the equation as follows:

$$F(x) = \frac{exp(x)}{(1 + exp(x))}.$$

Otherwise, the application of that model is similar to the OLS and LPM. Where the equation of the model is as follows:

$$P(y = 1 \mid x) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n) = F(\beta_0 + x\beta), \text{ given that:}$$

$$x\beta = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n. \text{ (Wooldridge, 2019)}$$

The model estimates the coefficients (β) using the Maximum Likelihood Estimation (MLE) method, which finds the set of coefficients that makes the observed outcomes most likely. Furthermore, the coefficients in the Logit model represent changes in the log odds of the outcome for one-unit changes in the predictor variables, offering a clear interpretation in terms of odds ratios. However, the interpretation of the coefficients is less straightforward than in the LPM, as they cannot be directly understood as changes in probabilities but as odds.

6. Results

The regression tables are in the Appendix due to space restrictions. The standard errors are in parenthesis and the significance levels indicated by the asterisks are as follows: * for $p < 0.1$, ** for $p < 0.05$, and *** for $p < 0.01$.

6.1 OLS - Linear Probabilistic Model

The table presents results, corrected for heteroscedasticity with the White Estimator due to binary dependent variable, of the Linear Probabilistic Model.

Results which are spotted across the models are that the Age of the employee has a consistently negative and statistically significant impact on the attrition. That implies that the older employees are less likely to leave the company, however, the impact is low where the probability of attrition increases by 0,3% with each additional year of the worker. Moreover, the logarithmic form of Distance from Home and the Dummy variable suggesting that the employees are in fact working over time have a significant and positive relationship with leaving attrition from the company, with 1% increase of a distance from home being a 2,4% increase in attrition and if a worker if in fact working overtime it leads to a 2% increase of probability of attrition. That entails that workers which live further from the office and the workers which are working over time are more likely to leave. Next, the logarithmic Monthly Income, a 1% increase in monthly income is associated with a decrease in the probability of

attrition by 6,2% to 10,2%, depending on the model, suggesting higher income reduces attrition risk.

Model 1 includes the basic independent variables which are serving as the baseline for comparison with more complex models. In Model 2 the Travel Frequency and the type of Department are introduced, and both of them are statistically significant. Firstly, frequent business travellers are more likely to leave, with a 1 unit increase corresponding to a 9,3% - 16,2%, model dependent heightened risk of attrition. While, being in the Sales department corresponds to a 7,3% increase. Model 3 expands further by adding variables like Rare Travel, being a Sales Representative or the Number of Companies Worked, among others. The most interesting variables, which are statistically significant are: Rare Travel increases attrition by 7,2%, being a Research and Development Scientist decreases by 8,5% and receiving Training decreases by 2,3%. Final Model (4), adds multiple variables which are related to the personal outlook on the job. High Environment Satisfaction (4,1% decrease), Job Involvement (6,7% decrease), Job Satisfaction (4% decrease), Relationship Satisfaction (2,2% decrease) and Work Life Balance (3,2% decrease), all of the estimations are statistically significant.

Model	Breusch-Pagan Test	Durbin-Watson Test	VIF (Max)
LPM 1	BP = 114.67, p < 2.2e-16	DW = 1.9295, p = 0.08829	1.472
LPM 2	BP = 123.26, p < 2.2e-16	DW = 1.9319, p = 0.09576	1.827
LPM 3	BP = 138.08, p < 2.2e-16	DW = 1.9193, p = 0.06064	2.400
LPM 4	BP = 147.47, p < 2.2e-16	DW = 1.8898, p = 0.01738	2.427

Table 2: Diagnostics of LPM

The diagnostic tests of the LPM models are presented in Table 2. The constantly significant Breusch-Pagan Test is expected in this model due to the binary dependent variable and the probabilistic approach. As previously mentioned the White estimator was used to correct the results of the models. The Durbin-Watson statistics are around the value of 2, with OLS 4 showing a p-value that suggests the presence of positive autocorrelation, which could affect the independence of the residuals. While the Variance Inflation Factor (VIF) values remain below the common threshold of 5 or 10, indicating that multicollinearity is not a

major concern, there is a slight upward trend in VIF values from OLS 1 through OLS 4, which warrants attention for potential multicollinearity issues as model complexity increases.

6.2 Logistic Regression - Logit

As in the linear probabilistic model the Age variable is negative across all Models, suggesting that older employees are less likely to quit their jobs, with decreasing the odds by ~2%, model dependent. Moreover, the Distance from Home and Overtime are positively correlated and statistically significant, in Model 1 the 1% increase in the log transformed distance indicates a 25,6% increase in the odds of attrition. And the effect becomes stronger in Model 4, with a 33.2% increase in attrition odds for each 1% increase. While for the Overtime variable, the odds are increasing by 147%, the value increases even further in the other models, what features the importance of lack of overtime on employee retention. A 1% increase in log-transformed monthly income reduces the odds of attrition by 117,2% in Model 4, which is highlighting income's significant negative effect on attrition likelihood. When it comes to the Job Role variable, the R&D Scientists show a significant decrease in attrition odds by 70.8% in Model 4, suggesting job role satisfaction or stability within specific roles. A 1% increase in the years since the last promotion increases the odds of attrition by 7.1% in Model 4, highlighting the importance of career progression in retention of employees. Each unit increase in the more personal variables such as environment satisfaction, job involvement, job satisfaction, relationship satisfaction, and work-life balance significantly reduces the odds of employee attrition by 40.5%, 60.7%, 39.4%, 22.2%, and 31.9% respectively. All of the mentioned impacts are statistically significant.

Model	% Correctly Predicted	McFadden R^2	R^2 Maximum Likelihood	R^2 Cragg & Uhler's
Logit 1	83.88%	0.165	0.135	0.230
Logit 2	84.15%	0.192	0.156	0.266
Logit 3	84.22%	0.192	0.156	0.266
Logit 4	84.56%	0.299	0.232	0.395

Table 3: Diagnostics of Logistic Regression

Table 3 presents the diagnostics of the Logistic Regression. As this models are built differently not the same metrics apply to assessment of them. It can be seen that the predictability correctness is increasing with a larger number of explanatory variables. Model 4 reaches a 84,56% correctly predicted attrition outcomes. The The McFadden R-squared value, which measures model fit compared to a null model, remains constant at 0.192 from Model 2 to Model 3, suggesting that there is no additional explanatory power gained. The "R-squared Maximum Likelihood" shows a slight increase, indicating a better fit as the models progress, although the change is marginal. Finally, Cragg & Uhler's R-squared, which aims to mimic the R-squared of a linear model, also remains constant at 0.266 from Model 2 to Model 3, suggesting that the proportion of explained variance in the dependent variable has not improved between these models.

7. Discussion

The complex relationship between retention of employees and other various factors that might be influencing that presents a multifaceted analysis. This paper focuses on finding statistically significant factors which are causing the attrition or lack of it in the companies based on the cross-sectional data.

The negative impact of Age on attrition in both models can be treated as evidence that older employees are more stable in there jobs, speaking generally. It might also imply that with the older age the threshold for change might be higher. That level might could be due to various factors such as increased responsibilities, financial stability needs, or lesser desire for career experimentation. The marginal increase in attrition probability or odds with each additional year, however, indicates that age alone is not the decisive factor of staying within the company. That also implies that organisations are compelled to create retention strategies.

The regression analyses depict the large negative effect of higher income on attrition likelihood, a finding that corroborates with Gardner et al. (2004) as highlighted by Das and Baruah (2013). Strategically and intuitively, the competitive compensation for employees is a particularly important factor. However, the extremely nuanced understanding of compensation's role, as discussed in the literature, points to the complexity of its impact. It suggests that while higher pay can be an effective tool in retention, its efficacy is dependent on the organisational culture within which it is implemented. Which is also highlighted in the

empirical findings of this paper. The positive relationship between overtime and attrition, which also aligns with Hyman and Summers (2004) regarding the negative implications of overtime work on employees' well-being. Furthermore, the significant decrease in attrition odds with increased job involvement and satisfaction also points out to the non-monetary incentives of employees, which was also found by Agarwal (1998) and Zhang (2016).

It is evident that there is no single factor that can be isolated when addressing attrition. Rather, a comprehensive approach that considers demographic, financial, environmental, and personal satisfaction elements is essential. When discussion the implications for organisations it is crucial to understand that there might be no solution that will fit all, as different demographic groups or job roles might have different needs. An overall strategy that arises from this paper is that compensation should not be only viewed as monetary. Tailoring retention strategies to address these varied needs is crucial for success.

8. Conclusions

This paper looks into the most influential factors for employee retention based on the cross-sectional data from IBM HR. The chosen methods for this analysis are Linear Probabilistic Model and the Logistic Regression. Age and Monthly Income emerge as influential factors, but on the other hand it is crucial to understand that solely those two factors are not responsible for lack of attrition. The compensation schemes should extend beyond the monetary incentives and a need for comprehensive approach of financial and personal incentives arises, which will acknowledge otherwise the employee contributions. The negative relationship between Work Life Balance, Job Satisfaction, Environmental Satisfaction, etc. with Attrition underlines that supportive work environment and non-monetary incentives are pivotal factors in retaining employees. Addressing these aspects through thorough and diverse retention strategies can help organisations mitigate attrition rates and create a more stable and productive workforce.

9. Limitations

It is essential to recognise some of the shortcoming of this paper. Firstly, this paper is based on a synthetic dataset from IBM and it may not fully capture the complexity and nuances of real-world employee behaviours and decisions. What makes this data potentially

non-generalisable into the real-world scenarios and actual organisational settings. Secondly, the OLS model's assumption of linearity and the potential for homoskedasticity issues, despite corrections, might not adequately model the non-linear relationships between employee attributes and attrition. While the Logistic regression is addressing some of these concerns, it still relies on assumptions about the distribution of independent variables and their relationship to the log odds of attrition, which may not hold in all cases. Thirdly, the focus on specific variables within the dataset due to the space of the paper means that potentially influential factors are not included in the dataset may have been overlooked. Lastly, attrition is a dynamic process influenced by time-dependent factors such as changing job markets, organisational changes, and personal life events, which a cross-sectional analysis cannot fully capture.

10. Further Research

To build up on findings and overcome the limitations of this paper there are few steps that can be taken. Firstly, acquiring data from organisations, which are not synthetic. Moreover, it would be good to acquire data from multiple organisations if the focus of the research would be an approach of real-world generalisable results. That collection of data would enable to perform an industry specific analysis or a entire labour markets analysis. Moreover, time bound data would also be a great additional expansion of the research to counteract the possible exogenous economic environment that was happening at that specific time. It would enable to more accurately account for changes in the workforce or organisational policies. Furthermore, including even more variables which may be influential in the modern labour economics such as remote work, freelance economy, or the increasing emphasis on mental health.

Going beyond the scope of this course's curriculum it also might be interesting to use data science's techniques such as various machine learning models and neural networks which might help uncover hidden patterns and interactions between variables that traditional econometric models may be unable to.

Bibliography

Agarwal, N. C. (1998). Reward systems: Emerging trends and issues. *Canadian Psychology*, *39*(1–2), 60–70. https://doi.org/10.1037/h0086795

Choughari, H. (2022, March 10). *How long should you stay in one job?* https://www.linkedin.com/pulse/how-long-should-you-stay-one-job-hassan-choughari/

Das, B. L., & Baruah, M. (2013). Employee Retention: A Review of Literature. *IOSR Journal of Business and Management*, *14*(2), 08–16. https://doi.org/10.9790/487x-1420816

Gardner, D. G., Van Dyne, L., & Pierce, J. L. (2004). The effects of pay level on organization-based self-esteem and performance: A field study. *Journal of Occupational and Organizational Psychology*, *77*(3), 307–322. https://doi.org/10.1348/0963179041752646

Hyman, J., & Summers, J. (2004). Lacking balance? *Personnel Review*, *33*(4), 418–429. https://doi.org/10.1108/00483480410539498

IBM Developer. (n.d.). https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/?mhsrc=ibmsearch_a&mhq=IBM%20HR%20Analytics%20Employee%20Attrition%20%26amp%3B%20Performance

IBM HR Analytics Employee Attrition & Performance. (2017, March 31). Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/code?datasetId=1067

Kane, G. C., Nanda, R., Phillips, A. N., & Copulsky, J. (2021). Redesigning the Post-Pandemic workplace. *MIT Sloan Management Review*, *62*(3), 12–14. https://www.scholars.northwestern.edu/en/publications/redesigning-the-post-pandemic-workplace

Linear Probability model. (n.d.). https://murraylax.org/rtutorials/linearprob.html

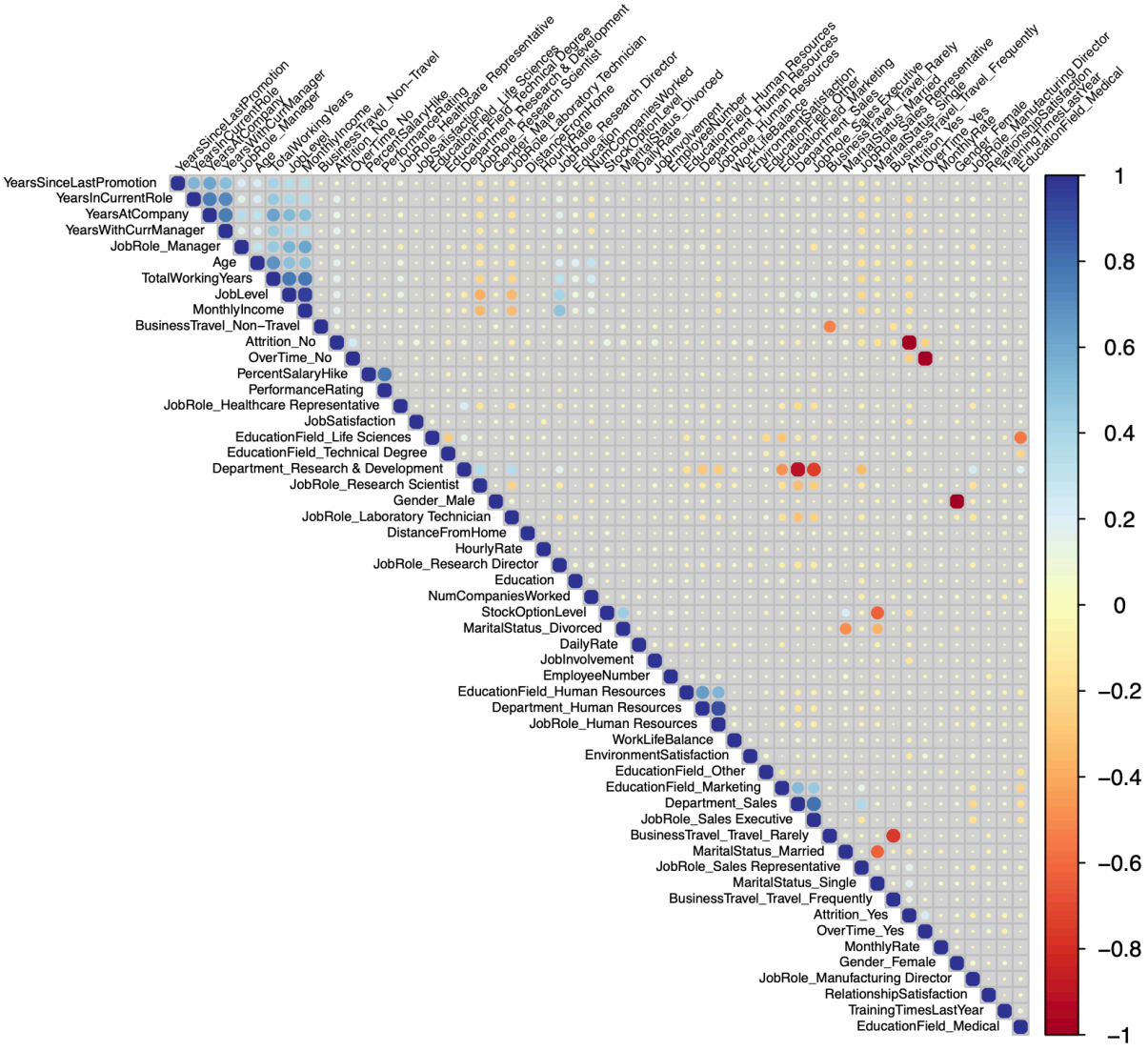
Pergamit, M. R., & Veum, J. R. (1999). What is a Promotion? *ILR Review*, *52*(4), 581–601. https://doi.org/10.1177/001979399905200405

Torra, V., Domingo-Ferrer, J., Mateo-Sanz, J. M., & Ng, M. K. (2006). Regression for ordinal variables without underlying continuous variables. *Information Sciences*, *176*(4), 465–474. https://doi.org/10.1016/j.ins.2005.07.007

Wooldridge, J. M. (2019). Introductory Econometrics: a Modern Approach. Cengage Learning.

Zhang, Y. (2016). A review of Employee turnover influence factor and Countermeasure. *Journal of Human Resource and Sustainability Studies*, *04*(02), 85–91. https://doi.org/10.4236/jhrss.2016.42010

Appendix



Correlation Chart

	Dependent variable:			
	(1)	(2)	(3)	(4)
Age	-0.003**	-0.002*	-0.003**	-0.003**
	(0.001)	(0.001)	(0.001)	(0.001)
Education	0.002	0.002	0.003	0.004
	(0.009)	(0.009)	(0.009)	(0.009)
log_DistanceFromHome	0.024***	0.024***	0.023***	0.024***
	(0.008)	(0.008)	(0.008)	(0.008)
OverTime_Yes	0.199***	0.198***	0.199***	0.209***
	(0.023)	(0.023)	(0.022)	(0.022)
log_MonthlyIncome	-0.062***	-0.082***	-0.095***	-0.102***
	(0.016)	(0.018)	(0.020)	(0.019)
log_PercentSalaryHike	-0.041	-0.035	-0.025	-0.054
	(0.039)	(0.038)	(0.038)	(0.055)
log_YearsAtCompany	-0.029***	-0.032***	-0.026***	-0.025***
	(0.008)	(0.008)	(0.008)	(0.008)
MaritalStatus_Single	0.115***	0.111***	0.087***	0.093***
	(0.021)	(0.020)	(0.027)	(0.026)
Travel_Freq		0.094***	0.156***	0.162***
		(0.025)	(0.031)	(0.029)
Department_Sales		0.073***		
		(0.020)		
Travel_Rare			0.072***	0.074***
			(0.023)	(0.022)
Gender_Female		-0.028	-0.029	-0.036**
		(0.018)	(0.018)	(0.017)
JobRole_Manager		0.049	0.057*	0.072**
		(0.031)	(0.032)	(0.032)

LPM Adjusted (part 1)

Job_Manufacturing_Dir			-0.052*	-0.043
			(0.027)	(0.027)
Job_Sales_Rep			0.110**	0.111**
			(0.053)	(0.051)
Job_RDScientist			-0.085***	-0.081***
			(0.027)	(0.026)
Job_Sales_Exec			0.036	0.040*
			(0.024)	(0.024)
log_YearsSinceLastPromotion		0.005	0.006*	0.007**
		(0.003)	(0.003)	(0.003)
log_NumCompaniesWorked			0.015***	0.015***
			(0.005)	(0.004)
StockOptionLevel			-0.023	-0.020
			(0.014)	(0.014)
log_TrainingTimesLastYear			-0.023***	-0.027***
			(0.009)	(0.008)
EnvironmentSatisfaction				-0.041***
				(0.008)
JobInvolvement				-0.067***
				(0.012)
JobSatisfaction				-0.040***
				(0.008)
PerformanceRating				0.016
				(0.033)
RelationshipSatisfaction				-0.022***
				(0.008)
WorkLifeBalance				-0.032**
				(0.013)
Constant	0.804***	0.921***	1.034***	1.661***
	(0.161)	(0.173)	(0.192)	(0.206)

LPM Adjusted (part 2)

WorkLifeBalance				-0.032**
				(0.013)
Constant	0.804***	0.921***	1.034***	1.661***
	(0.161)	(0.173)	(0.192)	(0.206)
==========	==========	==========	==========	==========
Note:			*p<0.1; **p<0.05; ***p<0.01	

LPM adjusted (part 3)

```
> stargazer(logit_model_1, logit_model_2, logit_model_3, logit_model_4, type = "text", font.size = "small")

=================================================================
                                Dependent variable:
                            ------------------------------------
                                      Attrition_Yes
                              (1)       (2)       (3)       (4)
-----------------------------------------------------------------
Age                        -0.020**  -0.019*  -0.028**  -0.026**
                            (0.010)   (0.010)   (0.011)   (0.011)

Education                             0.030     0.025     0.015
                                     (0.079)   (0.080)   (0.085)

log_DistanceFromHome       0.256***  0.265***  0.276***  0.332***
                            (0.077)   (0.079)   (0.081)   (0.086)

OverTime_Yes               1.482***  1.530***  1.599***  1.856***
                            (0.159)   (0.163)   (0.169)   (0.184)

log_MonthlyIncome          -0.672*** -0.853*** -0.981*** -1.172***
                            (0.151)   (0.174)   (0.206)   (0.222)

log_PercentSalaryHike       -0.348    -0.325    -0.188    -0.453
                            (0.345)   (0.352)   (0.360)   (0.554)

log_YearsAtCompany         -0.173*** -0.191*** -0.175*** -0.184***
                            (0.052)   (0.059)   (0.061)   (0.064)

MaritalStatus_Single       0.905***  0.895***  0.692***  0.781***
                            (0.159)   (0.163)   (0.224)   (0.236)

Travel_Freq                           0.743***  1.565***  1.852***
                                      (0.183)   (0.369)   (0.393)

Department_Sales                      0.701***
                                      (0.173)

Travel_Rare                                      0.855**  0.969***
                                                 (0.342)   (0.364)

Gender_Female                         -0.286*   -0.290*  -0.372**
                                      (0.164)   (0.168)   (0.177)
```

Logit Models (part 1)

JobRole_Manager	0.133	0.308	0.635	
	(0.528)	(0.565)	(0.589)	
Job_Manufacturing_Dir		-0.623	-0.484	
		(0.386)	(0.400)	
Job_Sales_Rep		0.550*	0.646**	
		(0.307)	(0.326)	
Job_RDScientist		-0.708***	-0.735***	
		(0.233)	(0.248)	
Job_Sales_Exec		0.465**	0.540**	
		(0.230)	(0.242)	
log_YearsSinceLastPromotion	0.038	0.058*	0.071**	
	(0.032)	(0.034)	(0.036)	
log_NumCompaniesWorked		0.167***	0.176***	
		(0.048)	(0.051)	
StockOptionLevel		-0.237*	-0.226	
		(0.136)	(0.142)	
log_TrainingTimesLastYear		-0.185***	-0.241***	
		(0.062)	(0.066)	
EnvironmentSatisfaction			-0.405***	
			(0.079)	
JobInvolvement			-0.607***	
			(0.119)	
JobSatisfaction			-0.394***	
			(0.078)	
PerformanceRating			0.036	
			(0.352)	
RelationshipSatisfaction			-0.222***	
			(0.079)	
WorkLifeBalance			-0.319***	
			(0.120)	

Logit Models (part 2)

WorkLifeBalance				-0.319***
				(0.120)
Constant	4.513***	5.568***	6.393***	13.378***
	(1.489)	(1.662)	(1.920)	(2.225)
Observations	1,470	1,470	1,470	1,470
Log Likelihood	-542.302	-524.616	-503.996	-455.408
Akaike Inf. Crit.	1,100.605	1,077.232	1,049.992	964.817
Note:				*p<0.1; **p<0.05; ***p<0.01

```
>
```

Logit Models (part 3)