Employee Attrition Analysis

Taewoo Kim

April 18, 2024

Contents

1. Introduction

According to the McKinsey & Company's article and their surveys (link) in 2021, forty percent of the employees responded that they are at least somewhat likely to quit in the next three to six months. Eighteen percent of the respondents said their intentions range from likely to almost certain. This survey was held across five countries (Australia, Canada, Singapore, the United Kingdom, and the United States) with broad range of industries. Although it was almost three years ago, coupled with the presence of generational disparities and challenging economic conditions, the answers are expected to remain similar now.

Employee attrition is defined as employees leaving their organizations for either unpredictable or uncontrollable reasons. While this term may carry negative connotations for companies, it doesn't always signify adverse conditions. In fact, it can serve as a positive indicator for healthy organizations, presenting opportunities to onboard new talent. This is particularly relevant in light of rapidly evolving trends and technological advancements.

However, if the employee attrition rate is excessively high, it could lead to problems such as unstable organizational culture, employee anxiety, lack of teamwork, increased training/hiring cost, etc. Generally, although it may highly depend on the industry, an appropriate attrition rate is considered 10%. In addition to that, even if a company keeps a good attrition rate, it could be also an issue if the majority of leavers were high performers. Therefore, employee attrition stands as a critical indicator that companies must closely monitor, given its potential long-term consequence stemming from the lack of attrition analysis.

2. Data Source

The data set, which was initially created by IBM data scientists in 2017, is from the Kaggle platform (link). As Human Resources data is highly confidential, the data provided here is fictional but structured very similarly to real data. It has 1470 rows and 35 columns, and each row indicates one employee data. This data set has no missing values. As for the response variable, Attrition, it consists of 237 of 'Yes' values and 1233 of 'No' values. The following table explains each variable with short description and type.

Variable Name	Role	Description	Dtype
Attrition	Response	Whether or not an employee has left the company, 'Yes' or 'No' response	object
Age	Predictor	Age of an employee	int64
BusinessTravel	Predictor	Employee travel frequency, 'Non-Travel', 'Travel_Rarely' or 'Travel_Frequently' response	object
DailyRate	Predictor	Salary level	int64
Department	Predictor	Department an employee belongs to, 'Human Resources', 'Research & Development' or 'Sales' response	object
DistanceFromHome	Predictor	The distance from work to home	int64
Education	Predictor	Education level (1: 'Below College', 2: 'College', 3: 'Bachelor', 4: 'Master', 5: 'Doctor')	int64
EducationField	Predictor	Education field an employee studied	object
EmployeeCount	Predictor	The number of employees for each row (all 1)	int64

EmployeeNumber	Predictor	Employee ID	int64
EnvironmentSatisfaction	Predictor	Indication of employee's environment satisfaction (1: 'Low', 2: 'Medium', 3: 'High', 4: 'Very High')	int64
Gender	Predictor	'Male' or 'Female' response	object
HourlyRate	Predictor	Hourly salary of an employee	int64
JobInvolvement	Predictor	1: 'Low', 2: 'Medium', 3: 'High', 4: 'Very High'	int64
JobLevel	Predictor	Level of a job (i.e., lower number indicates junior role)	int64
JobRole	Predictor	Position title of an employee	object
JobSatisfaction	Predictor	1: 'Low', 2: 'Medium', 3: 'High', 4: 'Very High'	int64
MaritalStatus	Predictor	'Divorced', 'Married' or 'Single' response	object
MonthlyIncome	Predictor	Monthly salary	int64
MonthlyRate	Predictor	Monthly salary rate	int64
NumCompaniesWorked	Predictor	Number of companies an employee worked before	int64
Over18	Predictor	'Y' or 'N' response	object
OverTime	Predictor	'Yes' or 'No' response	object
PercentSalaryHike	Predictor	The parentage of change in salary from the previous year	int64
PerformanceRating	Predictor	Performance rating, '3' or '4' response (i.e., higher number indicates higher performance)	int64
RelationshipSatisfaction	Predictor	1: 'Low', 2: 'Medium', 3: 'High', 4: 'Very High'	int64
StandardHours	Predictor	Standard working hours (all 80)	int64
StockOptionLevel	Predictor	How much company stocks an employee own from this company	int64
TotalWorkingYears	Predictor	Toal number of years of experiences of an employee	int64
TrainingTimesLastYear	Predictor	Employee's number of training times in last year	int64
WorkLifeBalance	Predictor	1: 'Bad', 2: 'Good', 3: 'Better', 4: 'Best'	int64
YearsAtCompany	Predictor	Toal number of years of working at the company	int64
YearsInCurrentRole	Predictor	Toal number of years of working in the current role	int64
YearsSinceLastPromotion	Predictor	Total number of years since the last promotion	int64
YearsWithCurrManager	Predictor	Toal number of years with the current manager	int64

Table 1. Variable Description

3. Problem Statement

Based on the dataset, the company's attrition rate is quite high (17.8%). Here, attrition rate is calculated as following:

{Number of employees that left during period ÷ Average number of employees for period × 100}

As mentioned earlier, this high attrition rate could potentially have adverse effects on the business, not only having to incur the training/hiring cost again but impacting its organizational culture.

Therefore, the aim of the analysis is to identify key contributing factors to attrition, facilitating consideration of preventative measures, and to develop machine learning models for future attrition prediction.

Additionally, assuming the causes of attrition could be different based on the department, generation and performance, the analysis will be conducted in three different ways to establish more concrete customized solutions.

Analysis based on these three aspects is expected to enable us to derive targeted solutions for each department, understand how different generations perceive work and identify the reasons why high performers leave the company. The following tables show the different attrition rates based on different groups. When it comes to the attrition rate by performance, although the rate is almost the same, it does not necessarily mean that the reasons for leaving the company are the same. Note that the definition of generation was referenced from Beresford Research (link) and calculated based on the year 2017.

Category	Attrition Rate
Human Resources	21.1%
Research & Development	15.1%
Sales	23.5%

Table 2. Employee Attrition Rate by Department

Category	Attrition Rate
High performers	18.1%
Average performers	17.8%

Table 3. Employee Attrition Rate by Performance

Category	Attrition Rate
Gen Z & Millennials (~ age 36)	23.7%
Gen X (age 37 ~ 52)	10.9%
Boomers (age 53 ~ 60)	13.1%

Table 4. Employee Attrition Rate by Generation

Category	P-value
Department	0.00448
Performance	0.91188
Generation	5.69E-07

Table 5. ANOVA Test Result

4. Methodology

As the analysis will be conducted across three different aspects, the following steps will be taken for each aspect.
First, before constructing predictive models, a comprehensive exploratory data analysis (EDA) phase will be undertaken. This can include checking potential outliers, correlations, distributions, and discovering underlying patterns, etc.

For instance, considering many of them are categorical variables, data preprocessing will be done by using one-hot-encoding. This will result in a larger number of variables, so feature selection methods will be employed to mitigate potential overfitting issues and manage the complexity of the model. In addition, some of the variables are different in scales. Thus, we will check the box plots and consider some scaling methods.

Following EDA, a variety of classification modeling will be employed. These include Logistic Regression, Random Forest, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), LightGBM, AdaBoost, and Multi-layer Perceptron Classifier (MLPClassifier). The model with the highest F1-score will be chosen as the best performer. Notably, as the data is imbalanced, resampling methods like SMOTE (Synthetic Minority Oversampling Technique) or SMOTE-NC will be applied prior to modeling.

For performance evaluation metrics, given the imbalanced nature of the data, the F1-score will be used for evaluation alongside the confusion matrix. To ensure the robustness of the models, cross-validation techniques will be utilized. The best model's performance will be compared with the performance of other models using T-Test and W-Test. After identifying the key indicators of attrition, possible preventative plans will be developed accordingly.

5. Analysis (Department)

5.1. Exploratory Data Exploration (EDA)

Firstly, before conducting EDA, the following variables were removed from consideration to simplify the problem:

1) Variables with no variance, implying that all values are the same (EmployeeCount, Over18, StandardHours).
2) Variables with similar implications, such as DailyRate, HourlyRate, and MonthlyRate (these were removed, while MonthlyIncome was retained).
3) Variables that no actionable insights can be derived (MaritalStatus, EmployeeNumber).

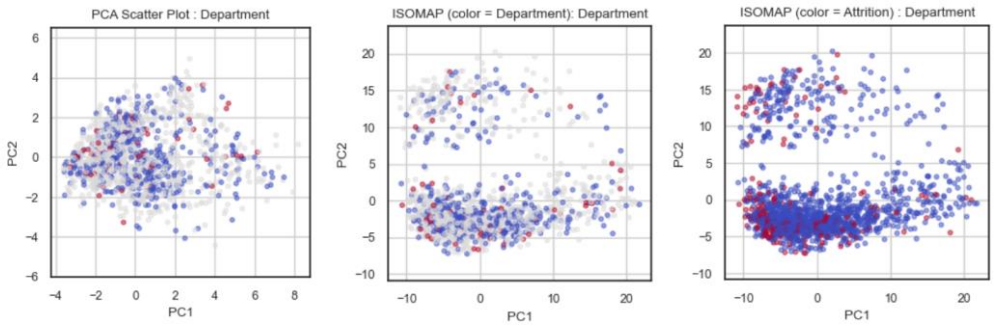

Figure 1. PCA, ISOMAP Plots

Examining PCA and ISOMAP plots could provide insights into the data distribution. In the plots above, different departments are represented by color, except for the ISOMAP plot on the right. From the visualizations, it appears that the data might be more non-linearly separable. However, it doesn't seem to be well-separated based on the response variable, Attrition. These plots suggest that departments are not distinctly separated from each other. Consequently, the reasons for attrition in each department may not differ significantly.

One point to consider is that the cumulative explained variance of the two principal components from PCA is only 0.23. This suggests that these components may not accurately represent the true distribution of the data.

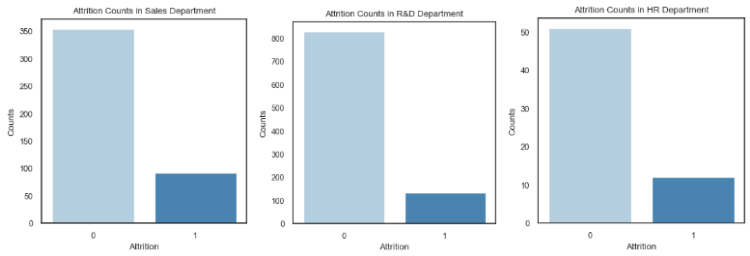

Figure 2. Attrition Count Plot Based on Each Department

Upon examining the target variable in three different departments, we can observe that the data is imbalanced. The attrition from Sales's department, 92, consist of nearly half of the total employee attrition, which stands at 237 across all departments.

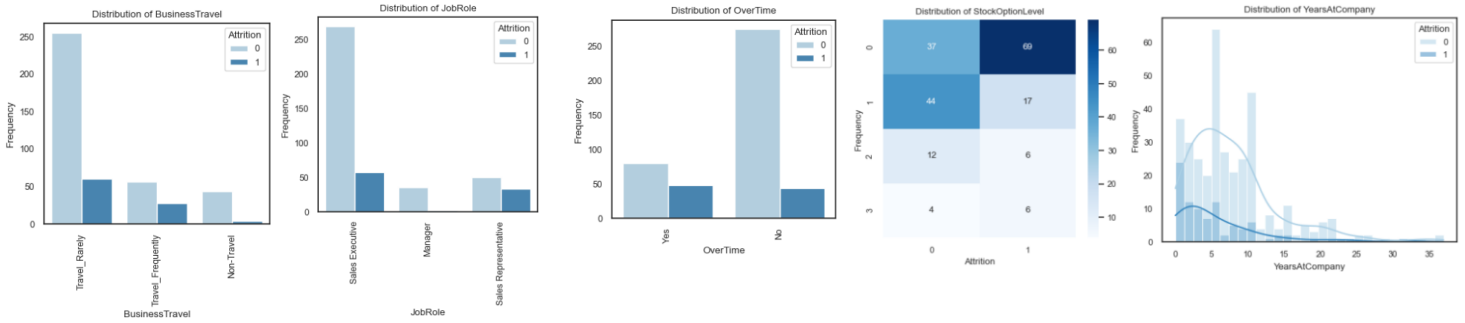

Figure 3. Bar plot, Heatmap and Distribution from Sales Department

The above plots are from the Sales department. We can observe that frequent business travel, the role of Sales Representative, working overtime, possessing fewer company stocks, and low monthly income could be contributing factors to attrition in the Sales department.

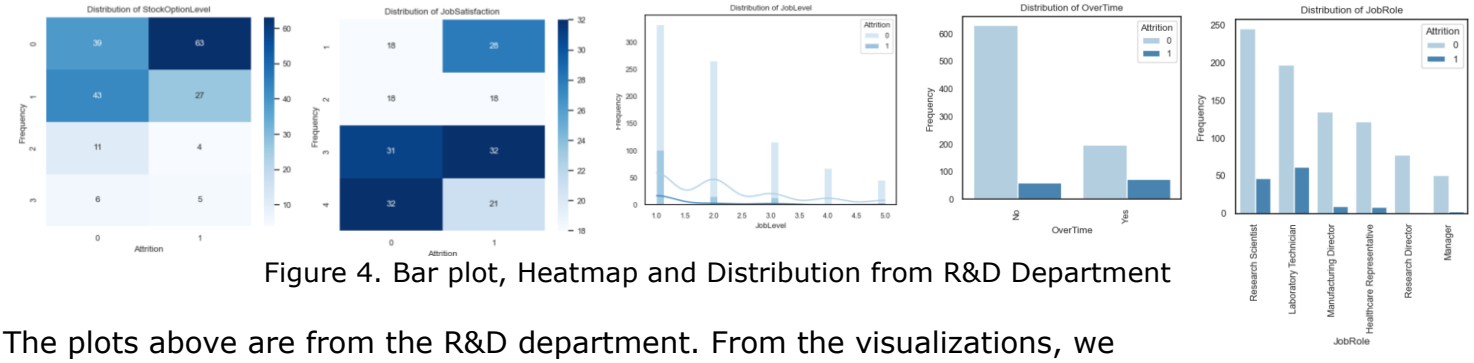

Figure 4. Bar plot, Heatmap and Distribution from R&D Department

The plots above are from the R&D department. From the visualizations, we can infer similar patterns that factors such as working overtime, possessing fewer company stocks, and having a lower monthly income (or lower job level) could potentially contribute to attrition. There are some different patterns observed in the R&D department as well. Job satisfaction level appears to have a stronger impact on attrition compared to the Sales department. Additionally, holding a Laboratory Technician role also seems to be a possible trigger for attrition.

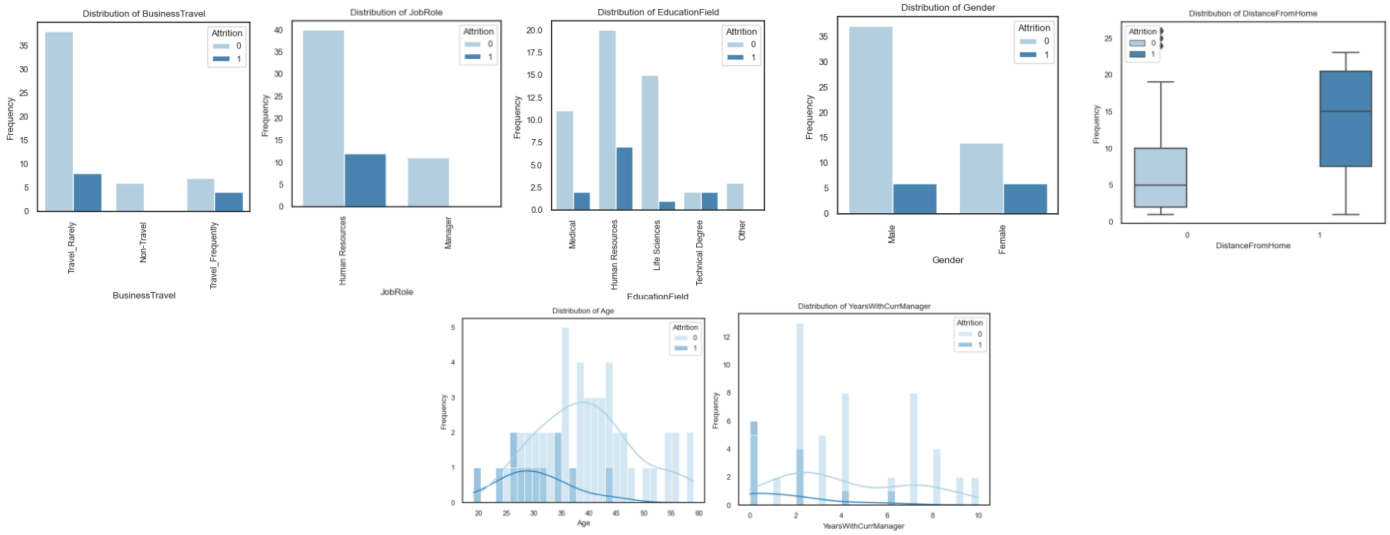

Figure 5. Bar plot, Box plot and Distribution from HR Department

The above plots are from HR department. There are more distinct patterns observed in the HR department, which could be attributed to the relatively smaller size of the data from this department. According to the results, factors such as frequent business travel, holding a role in Human Resources, majoring in Human Resources, being female, having a greater distance from home, belonging to younger age groups, and having fewer years with the current manager could potentially lead to attrition within the HR department.

After checking distributions and heatmaps, due to the presence of numerous categorical variables, certain categories have been combined with others to mitigate potential overfitting issues. For instance, the category of Doctor degree within the Education variable constitutes a very small portion of the data and has thus been merged with the Master degree category.

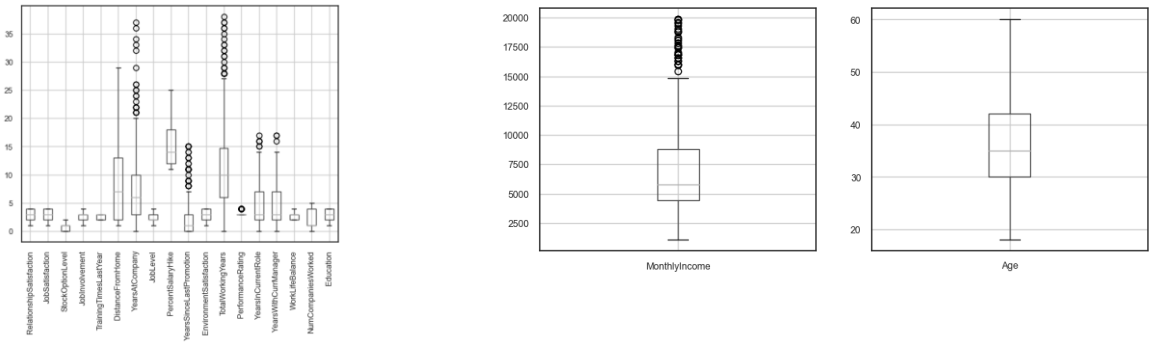

Figure 6. Box plots from Sales Department

The boxplots above are from the Sales department, but others show similar patterns. If we observe outliers in numerical variables, it's essential to address them appropriately. Assuming our employee data was correctly input and that the outliers are valid observations; we can transform the values of outliers to the maximum values of the non-outliers. This approach ensures that our models are not unduly influenced by extreme values while retaining the information conveyed by these outliers. It's worth noting that the PerformanceRating variable has only two unique values. Therefore, even though it may seem that the higher value is an

outlier, we will not remove it. Furthermore, it's evident from the wide range of scales among variables that scaling is necessary before modeling.

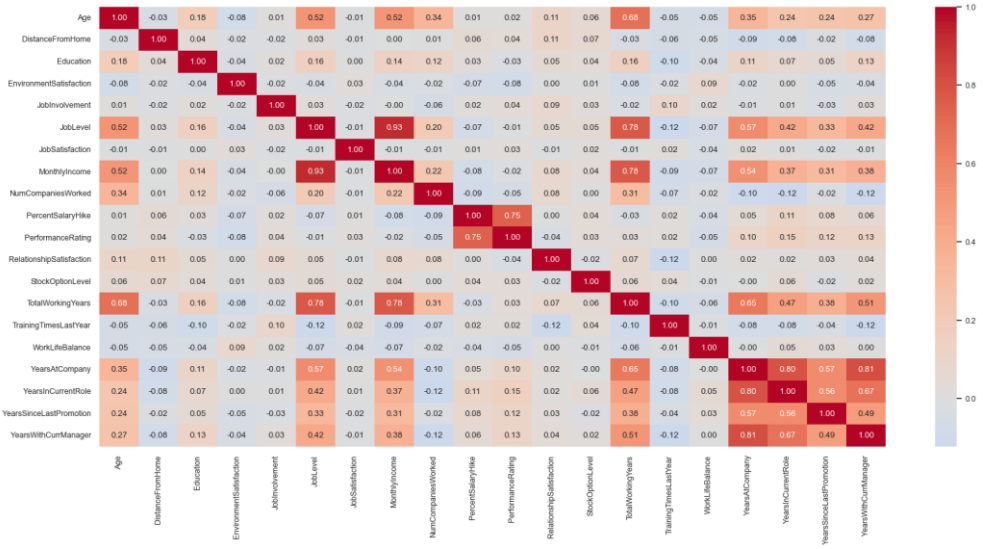

Figure 7. Correlation Matrix from Sales Department

Variable	GVIF	Df	GVIF^(1/(2*Df)
Age	2.095281	1	1.447509
BusinessTravel	1.132179	2	1.031523
DistanceFromHome	1.087487	1	1.042826
Education	1.120690	1	1.058626
EducationField	1.301817	4	1.033520
EnvironmentSatisfaction	1.061357	1	1.030222
Gender	1.037067	1	1.018365
JobInvolvement	1.050298	1	1.024840
JobLevel	9.664461	1	3.108772
JobRole	4.686324	2	1.471323
JobSatisfaction	1.041598	1	1.020587
MonthlyIncome	9.556177	1	3.091307
OverTime	1.088236	1	1.043185
PercentSalaryHike	2.415778	1	1.554277
PerformanceRating	2.408074	1	1.551797
RelationshipSatisfaction	1.095680	1	1.046747
StockOptionLevel	1.070159	1	1.034485
TotalWorkingYears	5.334525	1	2.309659
TrainingTimesLastYear	1.107807	1	1.052524
WorkLifeBalance	1.057462	1	1.028330
YearsAtCompany	6.411382	1	2.532071
YearsInCurrentRole	3.195010	1	1.787459
YearsSinceLastPromotion	1.670545	1	1.292496
YearsWithCurrManager	3.272589	1	1.809030

Table 6. VIF Result from Sales Department

After assessing correlations, the correlation matrix for numerical variables suggests a potential correlation between JobLevel and MonthlyIncome. The provided table displays the Variance

Inflation Factor (VIF) results obtained from R. Since our dataset includes categorical variables, we utilize the generalized VIF formula, calculated as GVIF^(1/(2*Df)). In this instance, a value of 3.16 is roughly equivalent to a VIF of 10. Hence, we can conclude that there is no significant multicollinearity issue among these predictor variables in the Sales department.

VIF Result are little different in R&D and HR departments, showing a bit more multicollinearity. The results from other departments can be found under the Appendix (Table 1 and Table 2).

5. 2. Feature Selection

Before conducting Feature Selection methods, the data was preprocessed by scaling, and dummy variables were created from categorical variables. Four different feature selection methods were employed: SelectKBest (Univariate selection method, using Mutual Information, Chi-Squared, ANOVA), LASSO, Forward Elimination, and Tree-based selection method (Random Forest Classifier). These methods were compared with the results obtained using all variables.

Optimal parameters were chosen using cross-validation, and Logistic Regression was utilized in 50 different training & test data splits to evaluate the performance of the model with optimal parameters. The following table shows the results. Notice that the data of the HR department is relatively small. Therefore, fewer variables were selected to avoid potential overfitting issues.

Department	Method	Selected # of features	F1 Score
Sales	SelectKBest	17	0.501015
	LASSO	22	0.575593
	Forward Elimination	17	0.484778
	RandomForestClassifier	23	0.568116
	Using all features	35	0.557547
R&D	SelectKBest	21	0.432894
	LASSO	25	0.451666
	Forward Elimination	18	0.401240
	RandomForestClassifier	25	0.461690
	Using all features	37	0.444757
HR	SelectKBest	8	0.495238
	LASSO	5	0.627047
	Forward Elimination	17	0.466809
	RandomForestClassifier	9	0.594000
	Using all features	34	0.453999

Table 7. Feature Selection Result

As a result, the selected variables from the best feature selection method will be used for future modeling. For example, the selected features from LASSO in the Sales department are as follows:

'Age', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction', 'JobInvolvement', 'JobSatisfaction', 'NumCompaniesWorked', 'PerformanceRating', 'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'WorkLifeBalance', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'BusinessTravel_Non-Travel', 'BusinessTravel_Travel_Frequently', 'EducationField_Marketing', 'Gender_Female', 'JobRole_Sales Representative', 'OverTime_No', 'OverTime_Yes'

We can observe that some of the features are those we previously identified meaningful while examining their distributions.

5. 3. Resampling

As observed earlier, the dataset is imbalanced. This imbalance can introduce biases in our future models, such as favoring the majority class or overfitting to the minority class due to insufficient data.

To address this issue, resampling methods were employed. Among these methods, SMOTE (Synthetic Minority Oversampling Technique) was selected as one of the oversampling techniques. SMOTE generates synthetic samples for the minority class to balance the dataset. Additionally, there is another type of SMOTE called SMOTE-NC, which is designed for datasets containing both numerical and categorical features. The results of SMOTE and SMOTE-NC were compared, and the method yielding better results was employed before every modeling step.

The following table shows the result. Cross-validation method such as GridSearchCV was utilized with 10 folds to find the best resampling ratio. Here, the resampling ratio equals to the number of samples in the minority class after resampling divided by the number of samples in the majority class.

Department	Method	Resampling ratio	F1 Score (cv=10)
Sales	SMOTE	0.55	0.629946
	SMOTE-NC	0.74	0.578453
R&D	SMOTE	0.80	0.496314
	SMOTE-NC	0.71	0.485419
HR	SMOTE	0.56	0.600000
	SMOTE-NC	N/A (no categorical features selected)	

Table 8. SMOTE and SMOTE-NC Result

5. 4. Modeling & Evaluation

8 different classification models were employed. These include Logistic Regression, Random Forest, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), LightGBM, AdaBoost, and Multi-layer Perceptron Classifier (MLPClassifier).

GridSearchCV was used with 10-fold cross-validation to identify the optimal combination of parameters for each model. While GridSearchCV can be computationally expensive due to the exhaustive search over the parameter grid, it ensures that the best hyperparameters are chosen to maximize model performance.

After obtaining the optimal parameters for each model, Monte Carlo cross-validation with 50 different splits was used to evaluate the performance of each model. The results are presented in the table below. Here, the base model refers to the best model's performance without any feature selection or resampling methods applied for comparison. The validation score and the test score are obtained from GridSearchCV using the optimal parameters, and the CV=50 score represents the mean of the F1 scores in 50 different splits using Monte Carlo cross-validation.

(Performance Metrics : F1-Score)

Dept.	Category	Logistic	Random Forest	LDA	SVM	XGB	Light GBM	Ada Boost	MLP	Base Model
Sales	Validation score (10 folds)	0.58681	0.49836	0.57347	0.60998	0.55484	0.52302	0.45233	0.59827	N/A
	Test score	0.68750	0.43243	0.66666	0.68571	0.51282	0.51851	0.30769	0.66666	
	CV=50 score	0.57046	0.45472	0.59209	0.60012	0.51753	0.55236	0.39573	0.59529	0.30902
R&D	Validation score (10 folds)	0.53053	0.53707	0.46764	0.50763	0.46732	0.50060	0.47626	0.51828	N/A
	Test score	0.58536	0.58536	0.51063	0.55814	0.61224	0.54902	0.47619	0.50000	
	CV=50 score	0.51361	0.49769	0.47660	0.50841	0.47934	0.50896	0.47186	0.45485	0.45276
HR	Validation score (10 folds)	0.61333	0.61333	0.54666	0.73333	0.30262	0.00000	0.48095	0.76666	N/A
	Test score	0.50000	0.33333	0.44444	0.33333	0.34782	0.00000	0.50000	0.57142	
	CV=50 score	0.63391	0.48179	0.61096	0.52486	0.31601	0.00000	0.51876	0.57840	0.43560

Table 9. Performance of Models

From the results presented above, the best classification model is highlighted in yellow. While SVM yielded the best performance for the Sales department, it utilized a Gaussian kernel. Therefore, for the interpretation of variable coefficients, the LDA model was selected.

The F1 scores of all the best models are above 0.5, indicating that the model performances are decent, though not exceptionally high. The following confusion matrices illustrate how the best models classified the employees.

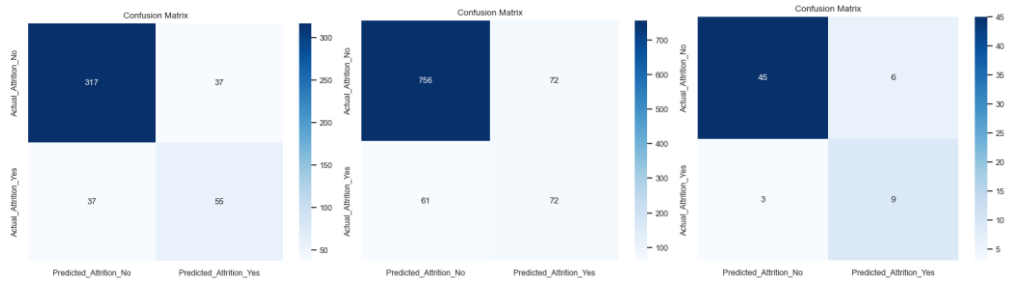

Figure 8. Confusion Matrices from The Best Models (Sales, R&D and HR in order)

Dept.	Category	Logistic	Random Forest	LDA	SVM	XGB	Light GBM	Ada Boost	MLP
Sales	T-Test	0.00874	0.18043	N/A	0.02314	0.85988	0.03217	0.74065	0.12861
	W-Test	2.56E-09	0.05060	N/A	1.86E-06	0.44075	0.00165	0.52102	0.00054
R&D	T-Test	N/A	0.32277	0.00874	0.71510	0.03299	0.76166	0.00550	0.00011
	W-Test	N/A	0.17452	2.56E-09	0.26651	0.01962	0.56734	0.00061	1.61E-12
HR	T-Test	N/A	0.00255	0.60922	0.01003	9.81E-16	4.09E-38	0.01026	0.20417
	W-Test	N/A	0.00034	0.39378	0.00220	4.96E-11	1.54E-09	0.00217	0.10247

Table 10. T-Test and W-Test P-value Result

The table above displays the results of the T-Test and W-Test. The T-Test is employed for parametric testing when the data is assumed to be normally distributed, while the W-Test is utilized for non-parametric testing when the data might not follow a normal distribution.

From the results, although the p-values of most of the models are significant, some exhibit p-values exceeding 0.05. This suggests that the performance of the best model could potentially be achieved by some models with different settings.

5. 5. Findings

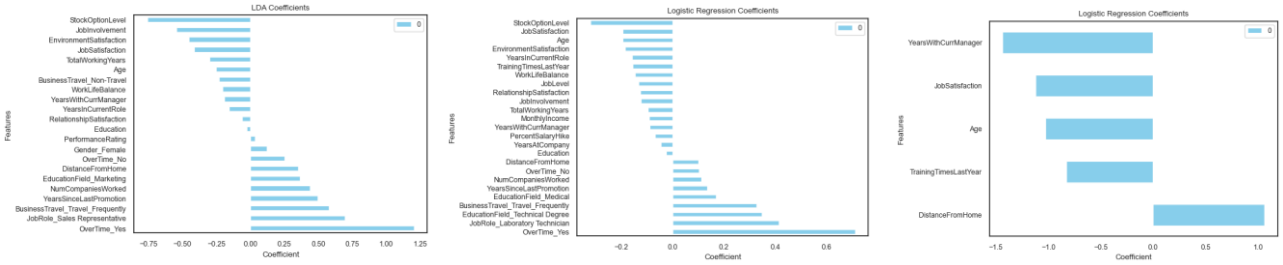

Figure 9. Degree of Coefficients from The Best Models

Dept.	Coefficient 1	Coefficient 2	Coefficient 3	Coefficient 4	Coefficient 5
Sales	OverTime_Yes	StockOptionLevel	JobRole_Sales Representative	BusinessTravel_ Travel_Frequently	JobInvolvement
	1.204696	-0.757403	0.694880	0.580845	-0.542451
R&D	OverTime_Yes	JobRole_Laboratory Technician	EducationField_Technical Degree	BusinessTravel_ Travel_Frequently	StockOptionLevel
	0.711534	0.414340	0.347567	0.327779	-0.320191
HR	YearsWith CurrManager	JobSatisfaction	Distance FromHome	Age	TrainingTimes LastYear
	-1.434815	-1.123067	1.063035	-1.025035	-0.828188

Table 11. Top 5 Coefficients and Variables

The coefficients displayed in the graph above reveal insights into the factors influencing attrition. Additionally, the table presents the top 5 coefficients and their corresponding variables. Across departments, overtime emerges as a significant contributor to attrition, particularly in the Sales and R&D departments. Conversely, in the HR department, employees with fewer years with their current managers are more prone to attrition. In response to these findings, the Sales department can focus on reducing overtime hours by streamlining inefficient processes and optimizing travel needs. Plus, interviewing some of the Sales Representatives could be beneficial in understanding the hidden aspects of attrition.

Similarly, the R&D department should address overtime issues and pay closer attention to the Laboratory Technician. Considering Laboratory Technician was not correlated with Technical Degree in correlation matrix (Appendix. Figure 1), employees with a technical degree in the R&D department may have more opportunities elsewhere, leading to more attrition.

In the HR department, it's evident that younger employees show higher rates of attrition, potentially linked to lower job satisfaction and shorter tenures with their current managers. The result between attrition and YearsWithCurrManager, which is linearly associated with

YearsAtCompany (as indicated in the Correlation Matrix in Appendix Figure 2), suggests that employees who have recently joined the company are more likely to leave. Therefore, the HR department should delve into the possible reasons for dissatisfaction among new employees, especially those of the younger generation, including commuting distance.

In addition, holding a lower amount of company stocks appears to influence employee attrition. To address this, one option could be to encourage employees to invest in company stocks or to offer company stocks instead of cash bonuses, providing employees with a stake in the company's success. This approach may help to reduce attrition by fostering a stronger sense of ownership and commitment among employees.

6. Analysis (Generation)

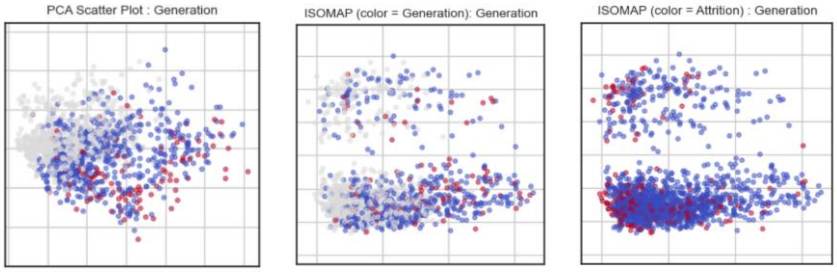

Figure 10. PCA, ISOMAP Plots for Different Generations

Same as the analysis for different departments, examining PCA and ISOMAP plots could provide insights into the data distribution. In the plots above, different generations are represented by color, except for the ISOMAP plot on the right. It doesn't seem to be well-separated based on the response variable, Attrition.

EDA, Feature Selection, SMOTE and modeling were employed in the same way as in the analysis for Department. Details can be found in the Appendix. The table below shows the results.

(Performance Metrics: F1-Score)

Gen.	Category	Logistic	Random Forest	LDA	SVM	XGB	Light GBM	Ada Boost	MLP	Base Model
Gen Z&M	Validation score (10 folds)	0.64735	0.56555	0.61736	0.63900	0.60434	0.61185	0.58129	0.64306	N/A
	Test score	0.57534	0.44117	0.60274	0.62162	0.55696	0.47272	0.44827	0.50000	
	CV=20 score	0.52463	0.55217	0.61424	0.61541	0.60114	0.52869	0.48866	0.59818	0.42887
GenX	Validation score (10 folds)	0.35833	0.35886	0.22690	0.35945	0.32690	0.21428	0.21547	0.34963	N/A
	Test score	0.24561	0.24000	0.23076	0.30188	0.16666	0.08695	0.09090	0.19512	
	CV=20 score	0.32816	0.29828	0.22872	0.32089	0.29828	0.19594	0.12302	0.28865	0.24181
Boomers	Validation score (5 folds)	0.15616	0.28000	0.13714	0.21490	0.21490	0.13333	0.19111	0.23000	N/A
	Test score	0.27586	0.22222	0.28571	0.22222	0.22222	0.22222	0.00000	0.15384	

	CV=5 score	0.20155	0.04000	0.08000	Failed Fitting	0.21756	0.00000	0.06666	0.29193	0.0877

Table 12. Performance of Models

The F1 scores of the best models, except for Gen Z&M, are below 0.5, suggesting that the model performances are poor. Particularly for GenX and Boomers, this indicates that attrition within those generations cannot be adequately explained using the given variables. Detailed confusion matrices can be found in the Appendix (Figure 7).

Gen.	Category	Logistic	Random Forest	LDA	SVM	XGB	Light GBM	Ada Boost	MLP
GenZ&M	T-Test	4.15E-05	0.00351	0.95508	N/A	0.47310	0.00030	1.46E-05	0.39910
	W-Test	0.00010	0.00070	0.86948	N/A	0.59581	0.00039	4.77E-05	0.10539
GenX	T-Test	N/A	0.17946	0.00613	0.74280	0.32330	4.96E-05	3.99E-09	0.08079
	W-Test	N/A	0.01068	0.00232	0.78412	0.40909	0.00010	1.90E-06	0.03623
Boomers	T-Test	0.03724	0.00097	0.03784	N/A	0.06387	9.42E-06	0.01493	N/A
	W-Test	0.06788	0.06250	0.12500	N/A	0.14412	0.06250	0.12500	N/A

Table 13. T-Test and W-Test P-value Result

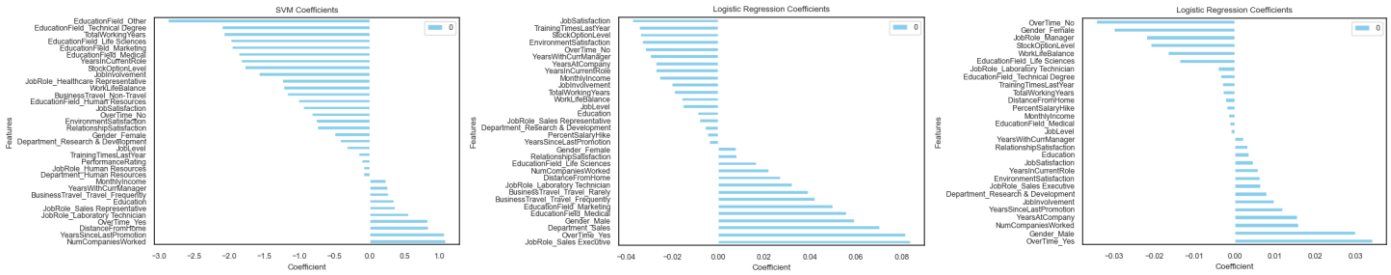

Figure 11. Degree of Coefficients from The Best Models

Gen.	Coefficient 1	Coefficient 2	Coefficient 3	Coefficient 4	Coefficient 5
GenZ &M	EducationField_Other	EducationField_Technical Degree	TotalWorkingYears	EducationField_Life Sciences	EducationField_Marketing
	-2.870457	-2.105032	-2.073723	-1.977515	-1.958403
GenX	JobRole_Sales Executive	OverTime_Yes	Department_Sales	Gender_Male	EducationField_Medical
	0.083852	0.081713	0.070388	0.059267	0.055849
Boomers	OverTime_Yes	OverTime_No	Gender_Male	Gender_Female	JobRole_Manager
	0.034259	-0.034259	0.029896	-0.029896	-0.021859

Table 14. Top 5 Coefficients and Variables

The table above presents the top 5 coefficients along with their corresponding variables. For Boomers, MLP cannot provide coefficients, hence Logistic Regression Coefficients are displayed instead.

When examining the Gen Z&M cohort, it's intriguing to note that EducationField shows strong negative coefficients overall, indicating that employees with backgrounds in those fields are less likely to leave their jobs. However, upon closer inspection of the coefficient chart (Figure 11), it becomes apparent that all categories in EducationField yield negative coefficients. This suggests

that employees with any educational background are less likely to leave work, which limits the insights we can glean and complicates interpretation. The negative coefficient associated with TotalWorkingYears implies that individuals with fewer years of work experience are more prone to leaving their jobs, even though Gen Z&M themselves inherently have fewer work experiences compared to other generations. When expanding our analysis to the top 10 coefficients, we can find negative coefficients of StockOptionLevel and JobInvolvement, which are more actionable items for the company to work on.

When considering Gen X and Boomers, although we can observe that some of the top 5 variables align with our earlier observations from distributions and heatmaps (see Figure 5, 6 in Appendix), the poor performance of the model renders the coefficient degrees less meaningful.

In summary, the analysis based on generations does not provide significant insights into why different generations leave their jobs, except for Gen Z&M. This suggests that while the younger generation may have specific reasons for attrition, such patterns are not discernible for Gen X and Boomers, indicating more random patterns or a lack of distinctive attrition patterns based on generation. Compared to the analysis based on departments, this implies that generational differences do not play a significant role in influencing attrition patterns.

7. Analysis (Performance)

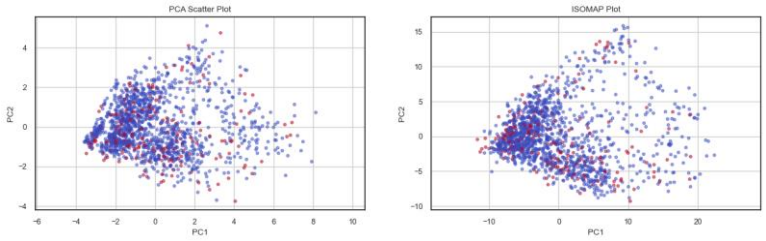

Figure 12. PCA, ISOMAP Plots for Different Generations

For the analysis based on performance, before delving into checking if the potential reasons for leaving work differ or not, let's first examine the differences between high performers and average performers in general.

As observed in the ANOVA Test presented in the introduction, the p-value suggests that there would be no significant difference between the two groups. Additionally, examining the PCA and ISOMAP plots reveals that they are distributed in very similar patterns. Figure 8 in the Appendix displays how the means of each variable are similar in both groups. Employing a clustering method such as KMeans Clustering further confirms that the distinct groups observed in the employees are primarily based on their age and factors related to years of experience, rather than performance. The results can be found in Figure 9 in the Appendix.

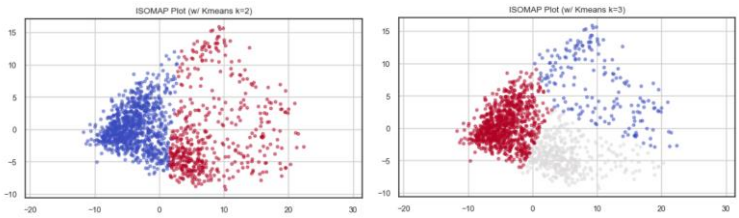

Figure 13. PCA, ISOMAP Plots for Different Generations (KMeans)

For simplicity, only Logistic Regression was used for this analysis using all variables and the results are as follows. The parameters were chosen using GridSearchCV. In the table, F1 score indicates the mean of F1 scores measured through cross-validation (cv=50).

Category	F1 Score (cv=50)
High performers	0.543144
Average performers	0.524187

Table 15. Logistic Regression Model Result

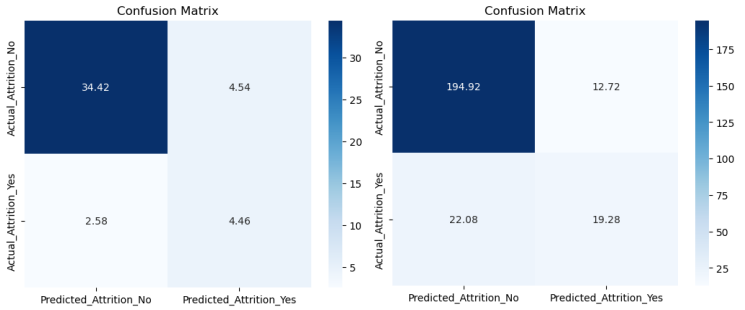

Figure 14. Confusion Matrices Result (means of 50 CV, High Performers: Left, Avg. Performers: Right)

Gen.	Coefficient 1	Coefficient 2	Coefficient 3	Coefficient 4	Coefficient 5
High Performers	OverTime_Yes	Distance FromHome	JobSatisfaction	EducationField_ Life Sciences	StockOptionLevel
	0.586759	0.443452	- 0.402694	0.379157	-0.355759
Avg. Performers	OverTime_Yes	BusinessTravel_ Travel_Frequently	Gender_Male	JobRole_Laboratory Technician	Department_Sales
	1.242454	0.739282	0.719595	0.718127	0.695348

Table 16. Top 5 Coefficients and Variables

Despite being a simple Logistic Regression model, the Logistic Regression yields a decent F1 Score (>0.5), effectively classifying attrition.

Both high performers and average performers show a tendency for attrition triggered by overtime work. However, while average performers are notably impacted by overtime (coefficient degree > 1), high performers appear to be less affected by overtime (coefficient degree < 1) but influenced by multiple factors simultaneously.

In comparison to average performers, high performers seem to be slightly more affected by job satisfaction (High Performers: -0.402694, Average Performers: -0.225797). Although the coefficients' degrees are not high, high performers also appear to be slightly more affected by distance from home (High Performers: 0.443452, Average Performers: 0.108800).

Therefore, to prevent attrition among high performers, the company could implement strategies to reduce overtime work and investigate potential issues that may lead to lower job satisfaction more thoroughly. Additionally, offering company stocks instead of cash bonuses could serve as an incentive to retain high-performing employees. Moreover, when hiring, considering the commuting distance of candidates could also be beneficial.

8. Appendix.

Variable	GVIF	Df	GVIF^(1/(2*Df)
Age	2.069437	1	1.438554
BusinessTravel	1.05969	2	1.0146
DistanceFromHome	1.03241	1	1.016076
Education	1.081094	1	1.039757
EducationField	1.090063	3	1.014476
EnvironmentSatisfaction	1.035732	1	1.017709
Gender	1.041036	1	1.020312
JobInvolvement	1.037176	1	1.018418
JobLevel	13.500953	1	3.674364
JobRole	9.210192	5	1.24861
JobSatisfaction	1.027223	1	1.01352
MonthlyIncome	17.075605	1	4.132264
OverTime	1.375775	1	1.172934
PercentSalaryHike	1.043584	1	1.021559
PerformanceRating	2.712502	1	1.646968
RelationshipSatisfaction	2.67853	1	1.636621
StockOptionLevel	1.034646	1	1.017175
TotalWorkingYears	1.033848	1	1.016783
TrainingTimesLastYear	5.035968	1	2.244096
WorkLifeBalance	1.049251	1	1.02433
YearsAtCompany	1.020778	1	1.010336
YearsInCurrentRole	7.038652	1	2.653046
YearsSinceLastPromotion	3.742468	1	1.934546
YearsWithCurrManager	1.66922	1	1.291983

Table 1. VIF Result from R&D Department

Variable	GVIF	Df	GVIF^(1/(2*Df)
Age	3.046399	1	1.745394
BusinessTravel	2.460624	2	1.252453
DistanceFromHome	2.016514	1	1.42004
Education	2.049866	1	1.431735
EducationField	8.344907	4	1.3037
EnvironmentSatisfaction	1.534444	1	1.238727
Gender	2.07651	1	1.44101
JobInvolvement	2.139608	1	1.46274
JobLevel	50.627545	1	7.115304
JobRole	10.977992	1	3.313305
JobSatisfaction	1.656485	1	1.287045
MonthlyIncome	69.765914	1	8.352599
OverTime	4.33626	1	2.082369
PercentSalaryHike	1.379279	1	1.174427
PerformanceRating	5.607418	1	2.367999
RelationshipSatisfaction	5.46967	1	2.338732
StockOptionLevel	1.893823	1	1.376162
TotalWorkingYears	1.632842	1	1.277827
TrainingTimesLastYear	17.240551	1	4.152174
WorkLifeBalance	1.627337	1	1.275671

YearsAtCompany	2.046973	1	1.430725
YearsInCurrentRole	11.300026	1	3.361551
YearsSinceLastPromotion	3.029324	1	1.740495
YearsWithCurrManager	1.916766	1	1.384473

Table 2. VIF Result from HR Department

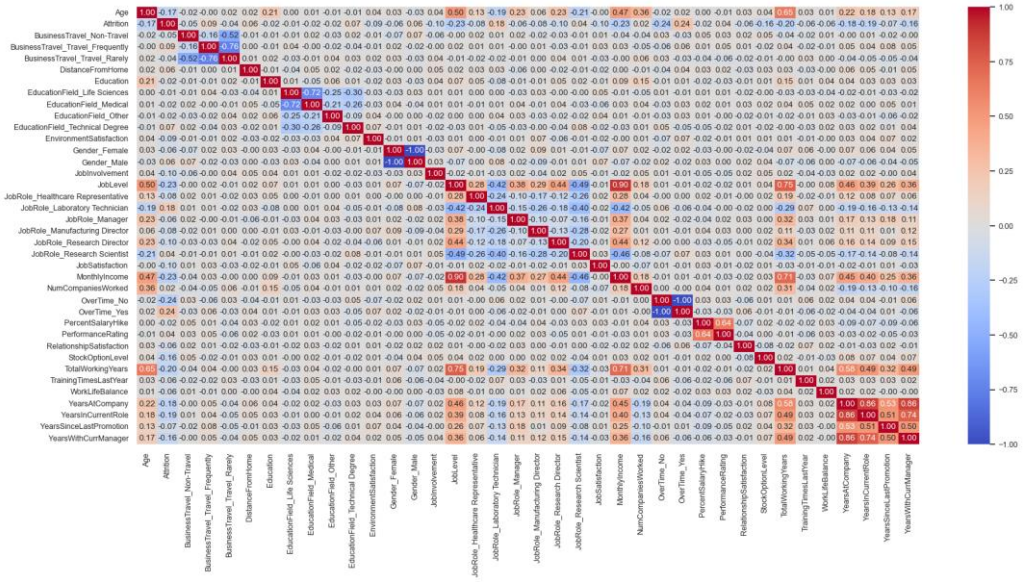

Figure 1. Correlation Matrix in R&D Department

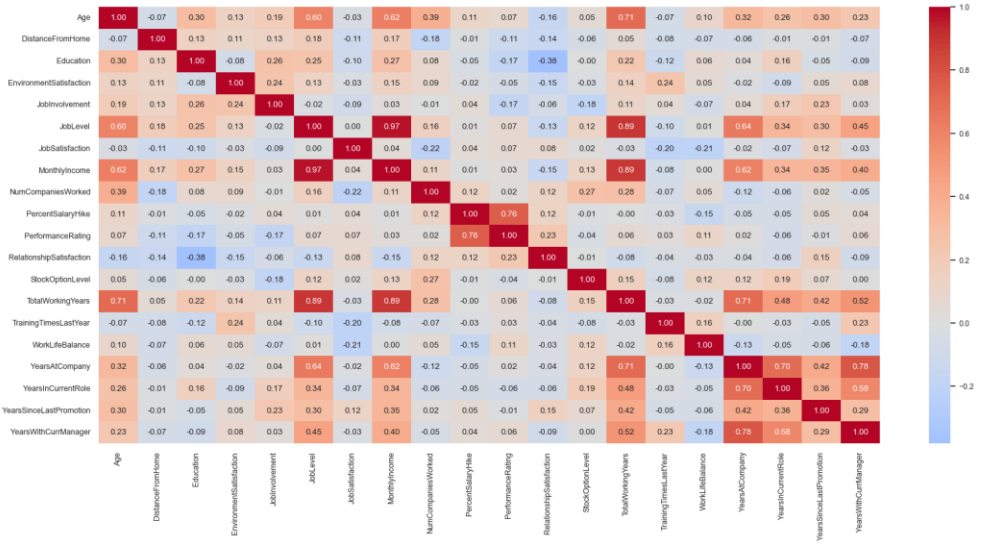

Figure 2. Correlation Matrix in HR Department

Variable	GVIF	Df	GVIF^(1/(2*Df))
BusinessTravel	1.05877	2	1.014379
Department	118.978599	1	10.907731
DistanceFromHome	1.059158	1	1.029154
Education	1.101649	1	1.049595
EducationField	1.548701	4	1.056199
EnvironmentSatisfaction	1.059104	1	1.029128

Gender	1.063923	1	1.031466
JobInvolvement	1.031167	1	1.015464
JobLevel	6.911726	1	2.629016
JobRole	584.852135	7	1.576326
JobSatisfaction	1.039092	1	1.019359
MonthlyIncome	5.646343	1	2.376203
NumCompaniesWorked	1.606099	1	1.26732
OverTime	1.052266	1	1.0258
PercentSalaryHike	2.577643	1	1.605504
PerformanceRating	2.56912	1	1.602847
RelationshipSatisfaction	1.033906	1	1.016812
StockOptionLevel	1.041198	1	1.020391
TotalWorkingYears	5.164528	1	2.27256
TrainingTimesLastYear	1.046678	1	1.023073
WorkLifeBalance	1.040156	1	1.01988
YearsAtCompany	9.525002	1	3.08626
YearsInCurrentRole	4.349091	1	2.085448
YearsSinceLastPromotion	1.525377	1	1.235062
YearsWithCurrManager	3.782953	1	1.944982

Table 3. VIF Result from GenZ&M

Variable	**GVIF**	**Df**	**GVIF^(1/(2*Df)**
BusinessTravel	1.186726	2	1.043729
Department	43.2067	2	2.563821
DistanceFromHome	1.080359	1	1.039403
Education	1.058339	1	1.028756
EducationField	2.976128	5	1.115232
EnvironmentSatisfaction	1.055722	1	1.027483
Gender	1.063446	1	1.031235
JobInvolvement	1.0781	1	1.038316
JobLevel	15.51679	1	3.939135
JobRole	366.117	8	1.446193
JobSatisfaction	1.059745	1	1.029439
MonthlyIncome	22.33023	1	4.725487
NumCompaniesWorked	1.229237	1	1.108709
OverTime	1.067858	1	1.033372
PercentSalaryHike	2.849122	1	1.687934
PerformanceRating	2.791081	1	1.670653
RelationshipSatisfaction	1.08981	1	1.04394
StockOptionLevel	1.05079	1	1.02508
TotalWorkingYears	3.191843	1	1.786573
TrainingTimesLastYear	1.061613	1	1.030346
WorkLifeBalance	1.088789	1	1.04345
YearsAtCompany	5.96871	1	2.443094
YearsInCurrentRole	2.880391	1	1.697172
YearsSinceLastPromotion	1.807171	1	1.344311
YearsWithCurrManager	3.316261	1	1.82106

Table 4. VIF Result from GenX

Variable	GVIF	Df	GVIF^(1/(2*Df)
BusinessTravel	2.192026	2	1.216778
Department	54.94322	2	2.722567
DistanceFromHome	1.730079	1	1.315325
Education	1.441785	1	1.200743
EducationField	35.6684	5	1.429646
EnvironmentSatisfaction	1.5216	1	1.233532
Gender	1.650228	1	1.284612
JobInvolvement	1.644669	1	1.282446
JobLevel	27.24723	1	5.219888
JobRole	1294.497	8	1.564971
JobSatisfaction	1.566505	1	1.251601
MonthlyIncome	40.00772	1	6.325165
NumCompaniesWorked	1.649924	1	1.284493
OverTime	1.534351	1	1.238689
PercentSalaryHike	3.094753	1	1.759191
PerformanceRating	3.502505	1	1.871498
RelationshipSatisfaction	1.496738	1	1.223412
StockOptionLevel	1.436453	1	1.198521
TotalWorkingYears	4.380944	1	2.093071
TrainingTimesLastYear	1.639995	1	1.280623
WorkLifeBalance	1.480792	1	1.216878
YearsAtCompany	8.836787	1	2.972673
YearsInCurrentRole	7.43337	1	2.726421
YearsSinceLastPromotion	2.835238	1	1.683816
YearsWithCurrManager	3.993273	1	1.998318

Table 5. VIF Result from Boomers

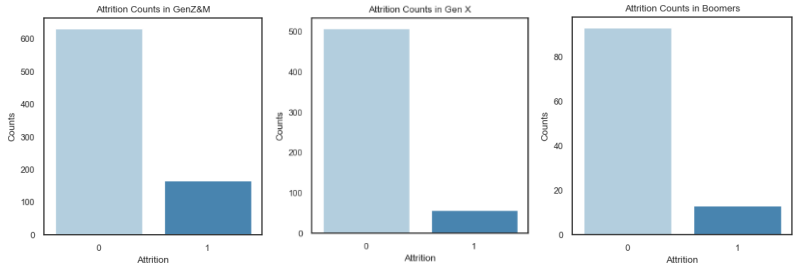

Figure 3. Attrition Count Plot Based on Each Generation

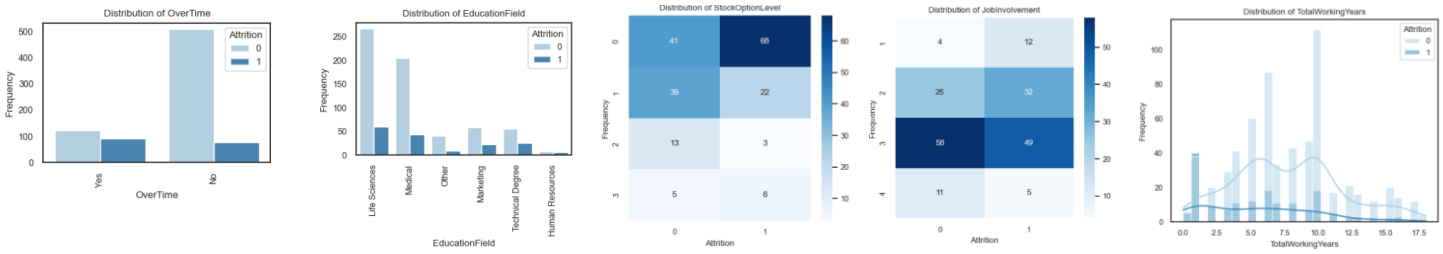

Figure 4. Bar plot, Heatmap and Distribution from Gen Z&M

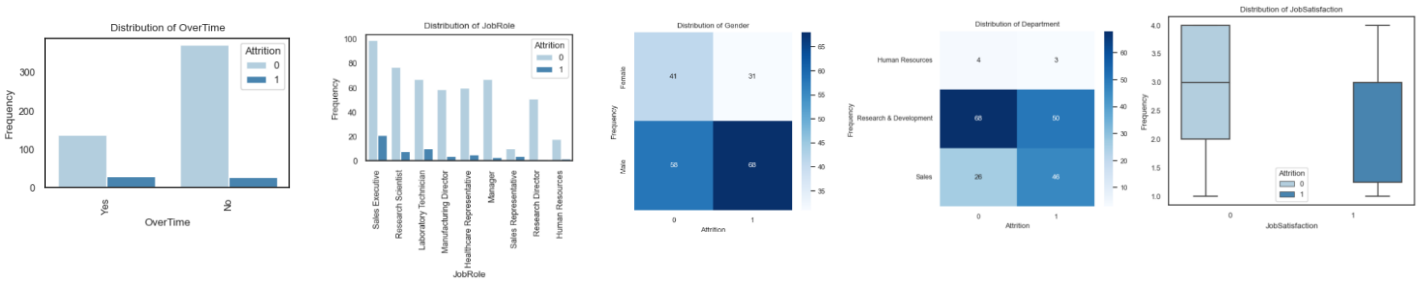

Figure 5. Bar plot, Heatmap and Distribution from Gen X

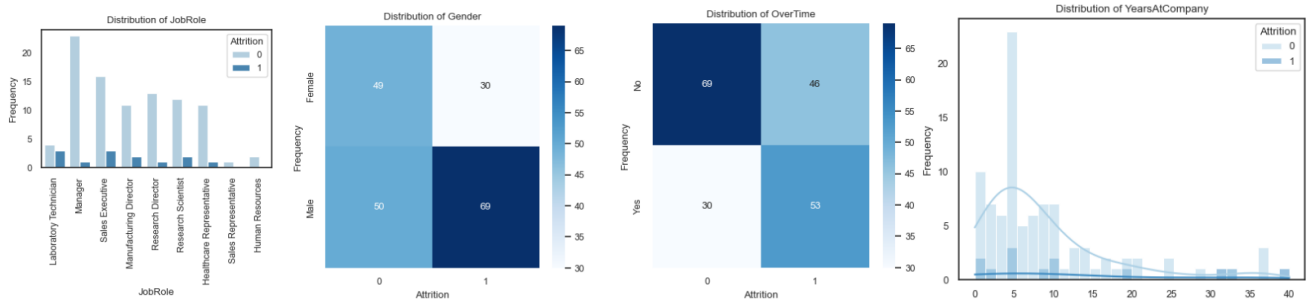

Figure 6. Bar plot, Heatmap and Distribution from Boomers

Generation	Method	Selected # of features	F1 Score
GenZ&M	SelectKBest	23	0.576922
	LASSO	34	0.618304
	Forward Elimination	22	0.585549
	RandomForestClassifier	33	0.615327
	Using all features	44	0.610161
GenX	SelectKBest	14	0.073879
	LASSO	18	0.175767
	Forward Elimination	22	0.162591
	RandomForestClassifier	32	0.192874
	Using all features	44	0.180736
Boomers	SelectKBest	1	0.000000
	LASSO	0	N/A
	Forward Elimination	22	0.076190
	RandomForestClassifier	29	0.090936
	Using all features	44	0.093587

Table 6. Feature Selection Result based on Generation (For Boomers, RF result was selected for a smaller number of variables)

Generation	Method	Resampling ratio	F1 Score (cv=10)
GenZ&M	SMOTE	0.37	0.646402
	SMOTE-NC	0.80	0.650563
GenX	SMOTE	0.45	0.306391
	SMOTE-NC	0.74	0.291833
Boomers	SMOTE	0.84	0.183333
	SMOTE-NC	0.70	0.207142

Table 7. SMOTE and SMOTE-NC Result based on Generation

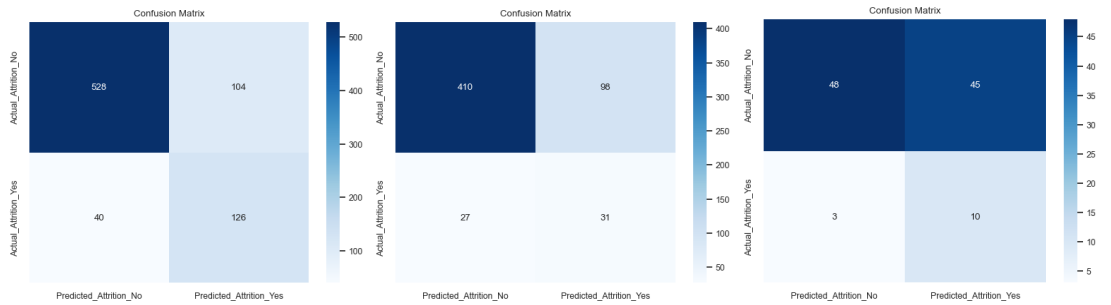

Figure 7. Confusion Matrices from The Best Models (GenZ&M, GenX and Boomers in order)

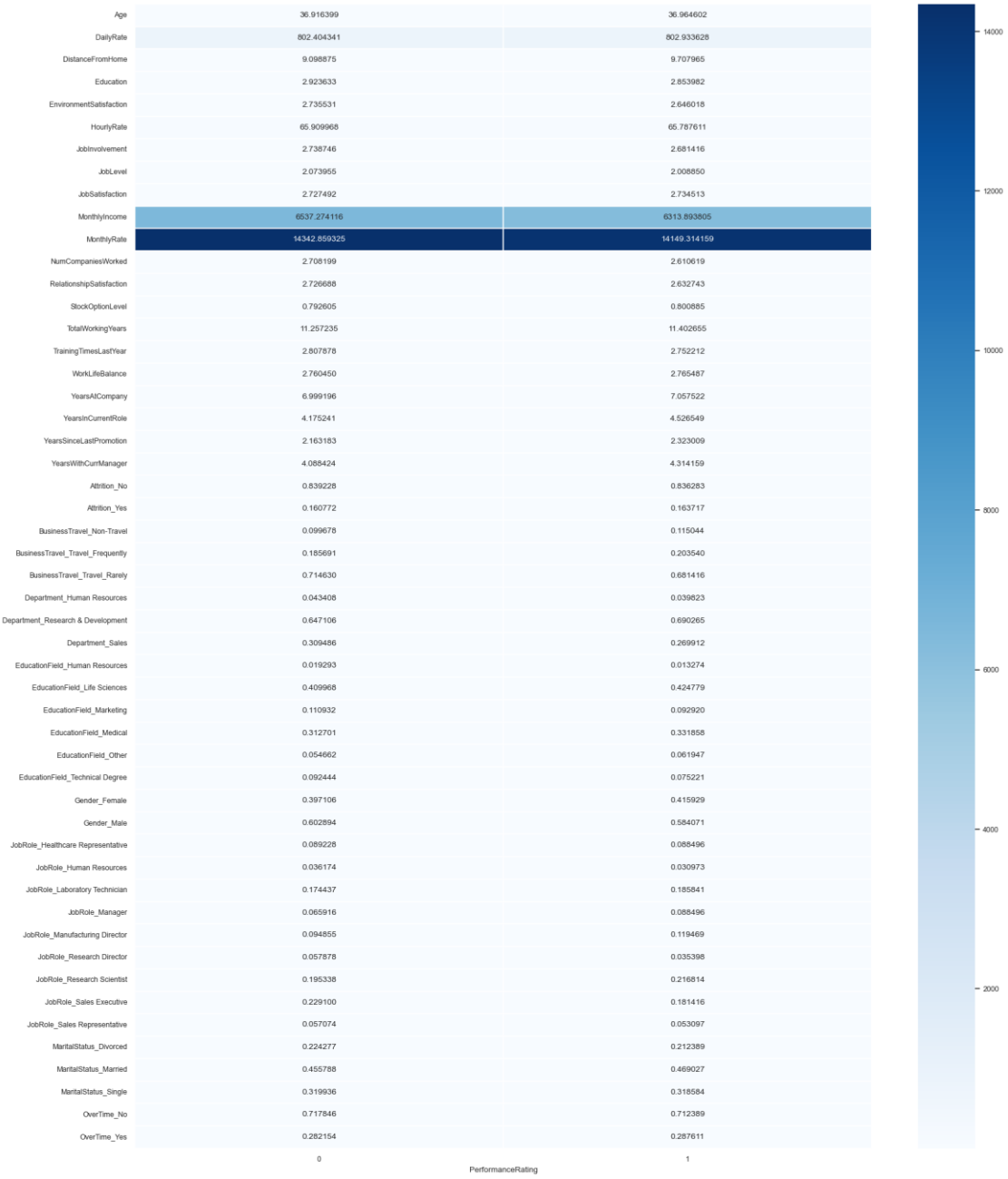

Figure 8. High Performers vs Average Performers Means Comparison

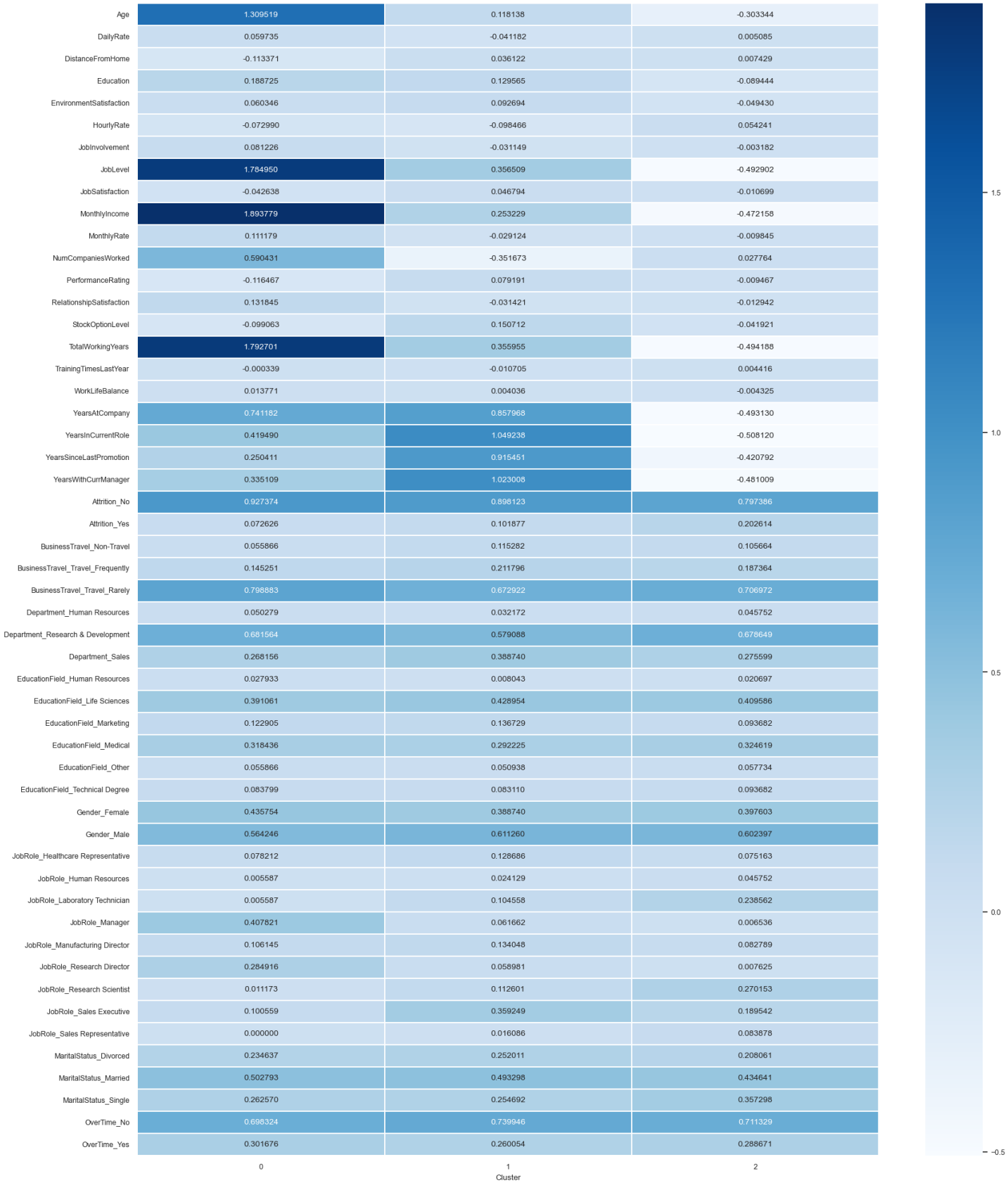

Figure 9. KMeans Clustering (k=3) Result (Cluster Means Comparison)