Employee Attrition Analysis

Taewoo Kim

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, 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 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 similarly to real data. It has 1470 rows and 35 columns, and each row indicates each 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'	object

		response	
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

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.

Assuming the causes of attrition could differ based on the department, we could examine different causes of attrition from each department. For simplicity, this analysis will focus on the department that shows the highest attrition rate, which is the Sales department. This approach will allow us to establish more concrete targeted solutions. The following tables show the different attrition rates based on different departments.

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

Upon further examination of the ANOVA result provided below, the significant p-value suggests that there is a meaningful difference in the means among the three different departments.

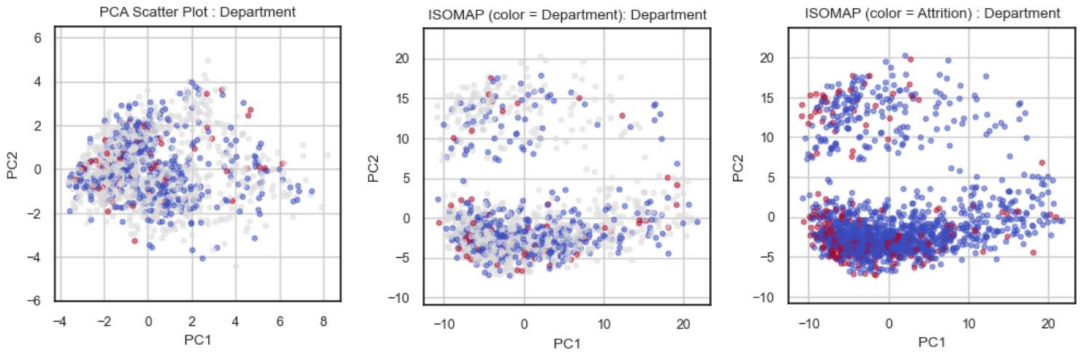
Category	Score
F-statistic	5.426843273
P-value	0.004485675

4. Exploratory Data Analysis

Firstly, before conducting Exploratory Data Analysis (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 indices, such as DailyRate, HourlyRate, and MonthlyRate (these were removed, while MonthlyIncome was retained).
3) Variables that no actionable insights can be derived (MaritalStatus, EmployeeNumber).

4.1. Distribution of Entire Data



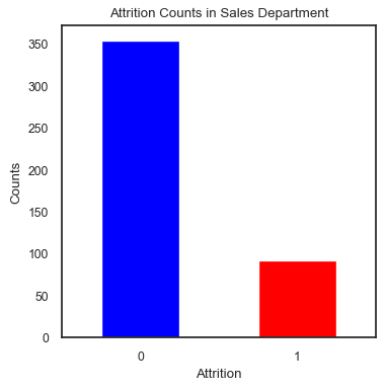
Before delving into the EDA of the Sales department, it might be beneficial to explore PCA and ISOMAP using the entire dataset. This approach could provide insights into the overall data distribution. In the plots above, different departments are represented by color (with blue indicating the Sales department), except for the ISOMAP plot on the right.

From the visualizations, it appears that the data might be more non-linearly separable. However, it

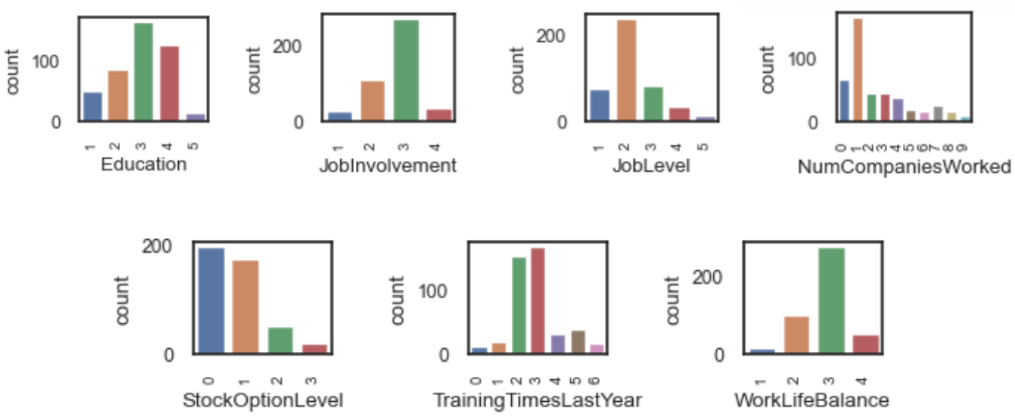
doesn't seem to be well-separated based on the response variable, Attrition. The plots suggest that departments are not distinctly separated from each other. Consequently, the reasons for attrition in each department may not differ significantly.

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.

4.2. Distribution of Sales Department Data

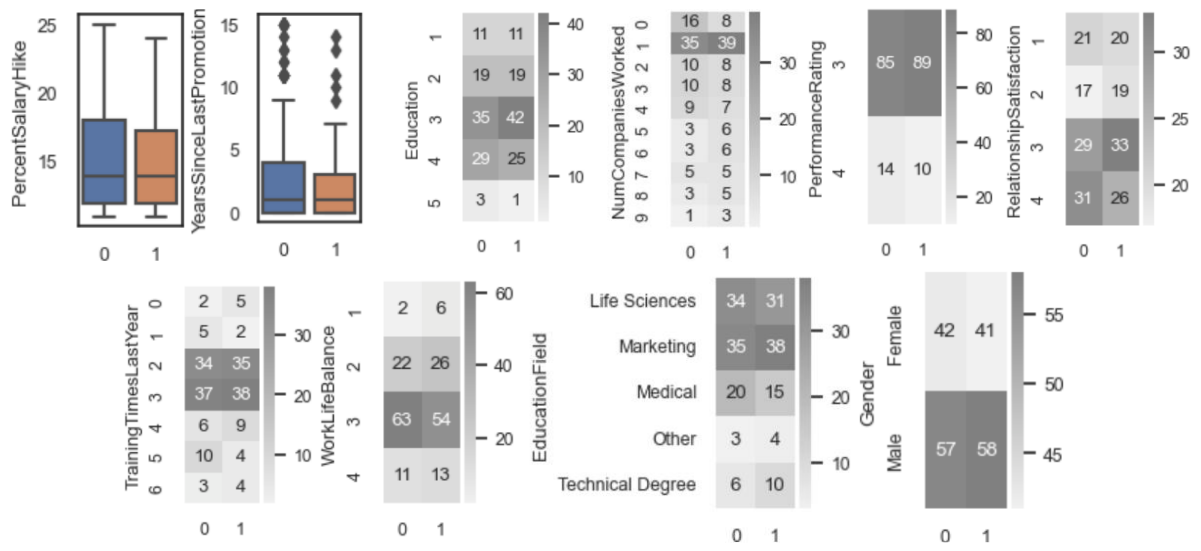


Firstly, when examining attrition within the Sales department, it's evident that 92 out of 446 employees, or approximately 20.6% (attrition rate is calculated with the average number of employees, therefore the number is slightly different from the attrition rate of 23.5%), have left the company during the given period. Notably, this accounts for nearly half of the total employee attrition, which stands at 237 across all departments.



Since there are numerous categorical variables, examining count plots can provide insights into the distribution of categories within each variable. From the plots, it's evident that certain categories within the variables occupy significantly smaller proportions compared to others. Consequently, it might be beneficial to combine these less represented categories within each variable.

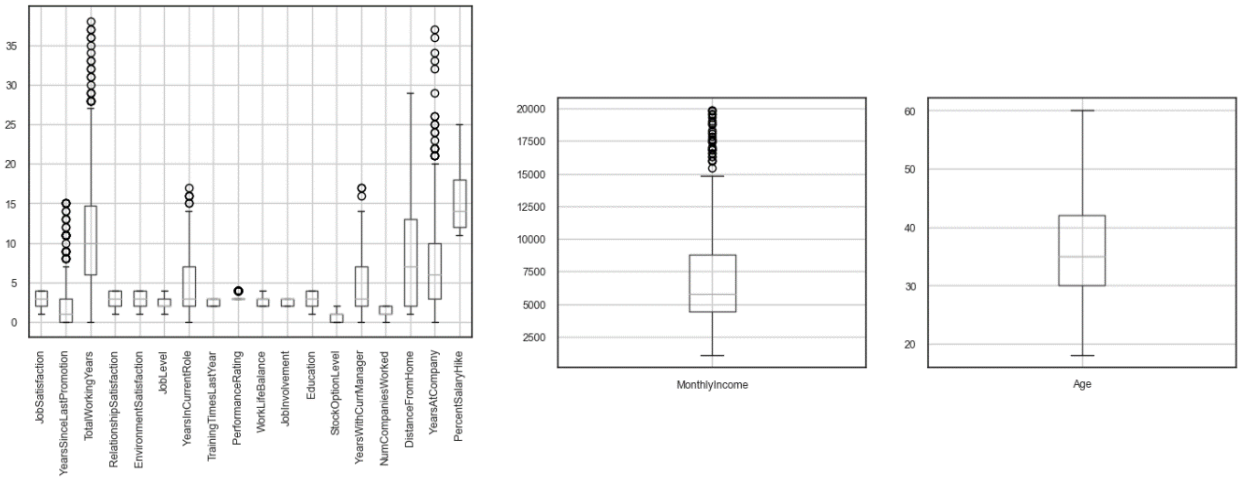
For instance, in the case of Job Involvement, categories 1 and 2 can be combined into a single category, and categories 3 and 4 can be combined. By doing so, when creating dummy variables later, we can mitigate the risk of overfitting by reducing the number of dummy variables generated.



Examining heatmaps and boxplots of categorical variables and some numerical variables with increasing order categories can offer insights into potential linear or monotonic relationships. From the analysis, it appears that the variables above exhibit minimal correlation with our response variable.

Considering this observation, we might opt to exclude these variables. However, for now, we will simply take note of this and proceed with feature selection. It's important to note that the numbers in the heatmaps represent the percentage of occurrences for each category (0 or 1).

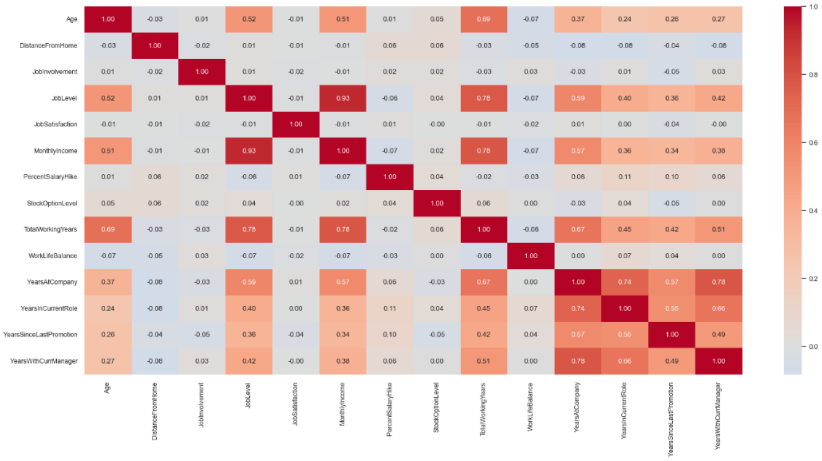
4.3. Outliers in Sales Department Data



If we observe outliers in numerical variables, it's essential to address them appropriately. Assuming the accuracy of our employee data 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.

Furthermore, it's evident from the wide range of scales among variables that scaling is necessary before modeling.

4.4. Multicollinearity in Sales Department Data



Variable	GVIF	Df	GVIF^(1/(2*Df))
Age	1.969494	1	1.403387
BusinessTravel	1.070447	2	1.017165
DistanceFromHome	1.044609	1	1.022061
JobInvolvement	1.020725	1	1.010309
JobLevel	8.417749	1	2.901336
JobSatisfaction	1.018245	1	1.009081
MonthlyIncome	8.556264	1	2.925109
OverTime	1.028287	1	1.014045
PercentSalaryHike	1.047702	1	1.023573
StockOptionLevel	1.039821	1	1.019716
TotalWorkingYears	4.652534	1	2.156973
YearsAtCompany	4.699538	1	2.167842
YearsInCurrentRole	2.53334	1	1.591647
YearsSinceLastPromotion	1.634514	1	1.278481
YearsWithCurrManager	2.792994	1	1.671225

After removing irrelevant variables and assessing multicollinearity, 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.

4.5. Feature Selection

Before conducting variable selection methods, the data was preprocessed by scaling, and dummy variables were created from categorical variables. Four different feature selection methods were employed: Univariate selection method (SelectKBest), LASSO, Backward Elimination, and Tree-based selection method (ExtraTreeClassifier).

For all four methods, optimal parameters were chosen using cross-validation, and Logistic Regression was utilized to evaluate the performance of the model with optimal parameters. The following are the results.

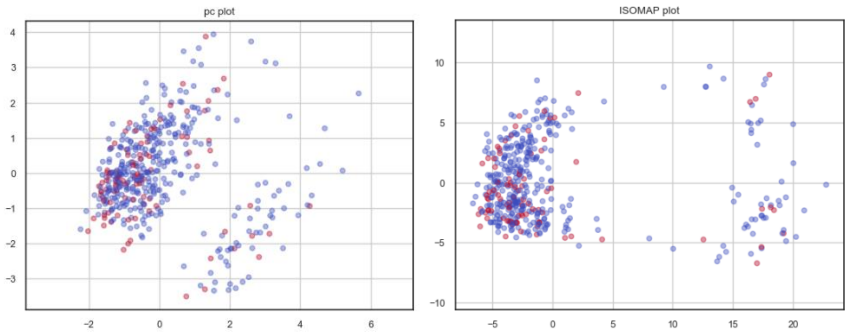
Method	Selected # of features (of 32)	ROC-AUC Score
SelectKBest	30	0.782301
LASSO	24	0.793093
Backward Elimination	16	0.798985
ExtraTreeClassifier	20	0.791424

As a result, we can proceed further with modeling using the selected features from the Backward Elimination method. The 16 features that were eliminated are as follows:

'BusinessTravel_Travel_Rarely', 'Education', 'EducationField_Life Sciences', 'EducationField_Marketing', 'EducationField_Medical', 'EducationField_Other', 'EducationField_Technical Degree', 'Gender_Female', 'JobLevel', 'MonthlyIncome', 'OverTime_No', 'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole'

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

5. Methodology



Before delving into the chosen methodologies, examining PCA and ISOMAP plots based on the selected variables from Backward Elimination could provide insights into the expected classification performance. In the provided plots, the red color indicates instances of attrition. From the visualization, while there seems to be some degree of separability, it's not distinctly separated by attrition. Notice that the explained variance is only around 28%, implying that it may not accurately represent the data. However, based on the pictures, we can anticipate that the performance of models might not be excellent.

The following methodologies were employed for this classification problem: LDA, QDA, Naïve Bayes, Logistic Regression, KNN, PCA-KNN, Random Forest, GBM (Gradient Boosting, and AdaBoost.

Here our baseline methods are LDA, QDA, Naïve Bayes, Logistic Regression, KNN and PCA-KNN.

Considering non-linear methodologies such as QDA, Naïve Bayes, KNN, and PCA-KNN do not provide coefficients information of the features, they might not be the best models for obtaining feature importance insights, but they could perform well in terms of predictive accuracy.

When it comes to parameter tuning, for KNN and PCA-KNN, odd numbers of 'k' ranging from 1 to 40 were explored, and the number of principal components was selected by examining the cumulative explained variance, aiming for a value of over 0.95.

For Random Forest, GBM, and AdaBoost, which have multiple hyperparameters like learning rate, max depth, number of trees, etc., GridSearchCV was utilized to identify the optimal combination of parameters for each model. Although GridSearchCV can be computationally expensive due to the exhaustive search over the parameter grid, it ensures that the best hyperparameters are chosen for maximizing model performance. After several adjustments to the parameter ranges, the final parameters chosen for each ensemble model are as follows:

Category	Parameter	GBM	AdaBoost	Random Forest
Input range	learning_rate	np.arange(0.01,0.05,0.01)	np.arange(0.01,0.05,0.01)	-
	n_estimators	[200, 500, 700]	[2, 5, 10, 20]	[20, 35, 50, 70]
	subsample	[0.1, 0.25, 0.5, 1]	-	-
	max_depth	[2, 3, 5, 10, 15]	-	[10, 13, 15, 20]
	min_samples_split	-	-	[2, 3, 4, 5, 7]
Optimal	learning_rate	0.03	0.02	-
	n_estimators	500	2	50
	subsample	0.25	-	-
	max_depth	2	-	10
	min_samples_leaf	3		1
	min_samples_split	-	-	2

Using cross-validation, 100 different training and test datasets were employed to train models with the optimal parameters. Model performance was assessed using the AUC-ROC metric, where a score close to 1 indicates a well-performing model, while a score close to 0.5 suggests a model with poor discriminatory power.

6. Result

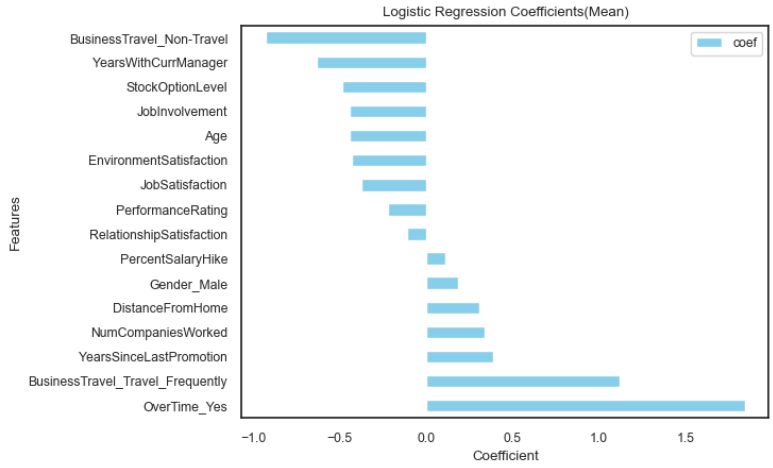
The results of each model's performance are as follows:

Category	LDA	QDA	Naïve Bayes	Logistic	KNN (k=33)	PCA-KNN (k=35)	Random Forest	GBM	AdaBoost
Training error (mean)	0.84003	0.91599	0.80007	0.84424	0.77812	0.75577	1	0.97346	0.66647
Test error (mean)	0.80944	0.70678	0.74385	0.80978	0.70599	0.68517	0.77359	0.78663	0.62943

From the table above, the Logistic Regression model exhibits the best performance in this dataset, with results nearly equivalent to those of LDA. Surprisingly, nonlinear classification methods like QDA, Naïve Bayes, KNN, and PCA-KNN did not fare well. This observation might suggest that attrition is more linearly separable within the dataset. Particularly, PCA-KNN performed even worse than KNN, suggesting that projecting the data onto principal components did not confer any significant advantage over using the original data, possibly due to the absence of multicollinearity in the dataset.

In comparison to baseline methods, Random Forest and GBM demonstrate stable performance, approaching that of Logistic Regression. However, the significantly high training error suggests potential overfitting. Despite the use of GridSearchCV to explore various parameter combinations,

there may still be room for parameter adjustment to achieve better fitting. If overfitting concerns were addressed, Random Forest and GBM could potentially yield the best results. Conversely, AdaBoost did not exhibit overfitting issues but yielded the poorest performance among all models.



The plot above shows the mean coefficient of each feature across 100 rounds of cross-validation. According to the results, working overtime appears to be the most significant factor contributing to employee attrition in the Sales department, followed by the frequency of business travel and the shorter tenure with a current manager. While not as influential, factors such as company stock options, higher job involvement, older age, and greater environmental satisfaction seem to contribute to employees staying in the company.

7. Findings

Although the performance of our best classification model is not significantly great, it provides insights into potential reasons why employees in the Sales department leave the company. Given the nature of work in the Sales department, which often entails frequent business travel and overtime hours, it is reasonable to consider these factors as contributing to attrition.

To address this, the Sales department could optimize travel routes to reduce travel frequencies and time, thereby potentially reducing employee attrition. Additionally, they could investigate the main factors driving overtime hours and identify actionable items to mitigate this issue.

Regarding model performance, it's noteworthy that ensemble methods such as Random Forest and Gradient Boosting performed almost as well as the best model, despite indications of overfitting. By fine-tuning the parameters of these models, we may be able to develop more robust models for predicting future employee attrition.

8. Appendix

Course_project_v0-Copy1.ipynb