Employee and HR Analytics: A Predictive Modeling Approach on Attrition

BANA 273: Machine Learning Analytics

Professor Mingdi Xin

Team 6A

Philip Tong, Cody Asano,

Xinru Zhao, Pratiksha Gund, Shin Yang

UC Irvine MSBA

December 6, 2024

Table of contents

0. Executive Summary

Employee turnover causes significant challenges to organizations, leading to increased recruiting costs, loss of knowledge, and disruptions of team dynamics. This report analyzes the IBM HR Analytics Employee Attrition & Performance dataset to identify key factors contributing to employee attrition and provide actionable recommendations to mitigate turnover. The dataset includes 1,470 employee data with 35 attributes including environmental satisfaction, salary, hourly rate, job role, job satisfaction, performance ratio, marital status, overtime, etc. Using advanced machine learning models including Random Forest, Logistic Regression, and K Nearest Neighbors, we achieved the highest accuracy of 0.87 from the Random Forest classifier, which also highlighted key features driving attrition.

Based on these findings, organizations are advised to prioritize improving work-life balance by reducing overtime work and tailoring compensation strategies to retain high-performing employees. By proactively addressing these factors, HR departments can enhance employee experience and reduce turnover rates, resulting in improved organizational stability and cost savings.

1. Introduction

Employee retention is a critical challenge for organizations, as high turnover rates can disrupt operations, increase recruitment and training costs, and negatively impact team dynamics and morale. Addressing this issue is vital for maintaining a stable workforce and ensuring organizational success. The business idea underpinning this analysis focuses on leveraging data-driven insights to proactively identify at-risk employees and implement effective retention strategies. This study utilizes the IBM HR Analytics Employee Attrition & Performance dataset, which includes detailed information related to employee retention across 35 attributes. These attributes span various dimensions such as environmental satisfaction, compensation, and job performance ratings with an “Attrition” variable, indicating whether an employee has left the company, making it an ideal resource for analyzing turnover trends. The importance of this project lies in its potential to address a pressing business problem by combining advanced analytical techniques with real-world HR data. Organizations today face increasing competition for talent, and understanding the drivers of employee attrition is crucial for retaining top performers and maintaining a competitive edge. Through the application of machine learning models such as Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN), this analysis aims to build a robust predictive framework for identifying attrition risk. These methods allow us to uncover key patterns in employee behavior, identify influential factors contributing to turnover, and offer actionable recommendations to help HR departments enhance employee retention and overall organizational stability.

2. Data

2.1 Acquisition

The HR Analytics dataset used in this analysis is sourced from Kaggle and is titled *IBM HR Analytics Employee Attrition & Performance*[1]. It has 1470 rows and 35 columns without any duplicate rows or missing values. The dataset captures various factors that may contribute to employee attrition, providing valuable insights to help companies reduce recruitment costs, retain institutional knowledge, maintain team cohesion, and improve overall workforce management. By understanding these factors, organizations can develop targeted strategies to enhance employee satisfaction and minimize turnover rates.

2.2 Description

Below is a detailed description of each of the features in the dataset:

Variables	Description
Age	Age of each employee.
Attrition	Reduction of Workforce. Yes: quit / No: not quit
BusinessTravel	Travels during business. None, Rarely, Frequently
DailyRate	Compensation or wage for an employee per day
Department	Research & Development, Sales, Human Resources
DistanceFromHome	Distance from home to work
Education	Employee's education level measured by 1-5
EducationField	Employee's education Field. Life Sciences, Medical, Marketing, Technical degree, Other.
EmployeeNumber	The specific number of an employee
EnvironmentSatisfaction	It is measured by 1-4
Gender	Male, Female
HourlyRate	Compensation or wage for an employee per hour
JobInvolvement	Job Involvement is measured by 1-4
JobLevel	Job level is measured by 1-5
JobRole	Job role for employee.
JobSatisfaction	JobSatisfaction is measured by 1-4
MaritalStatus	Married, single, divorced
MonthlyIncome	Monthly Income for employees
MonthlyRate	Compensation or wage for an employee per month
NumCompaniesWorked	Number of Companies Worked
OverTime	Whether an employee works overtime. Yes: overtime / No: not overtime

[1] *IBM HR Analytics Employee Attrition & Performance* https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

PercentSalaryHike	the percentage increase in an employee's salary
PerformanceRating	The rating score of performance with 3 and 4
RelationshipSatisfaction	The relationship satisfaction with 1-4
StockOptionLevel	the level or tier of stock options granted to an employee with
TotalWorkingYears	The total working years for an employee
TrainingTimesLastYear	The Training times an employee had Last Year with 0-6
WorkLifeBalance	The Work Life Balance for an employee with 1-4
YearsAtCompany	The time an employee stays at the company
YearsInCurrentRole	The time an employee working in their current role
YearsSinceLastPromotion	The time after promoted
YearsWithCurrManager	The time an employee worked with current manager

There are 9 categorical variables (“Attrition”, “Business Travel”, “Department”, “EducationField”, “Gender”, “JobRole”, “MaritalStatus”, “Over18”, “OverTime”) and 26 numerical variables, including 3 constant variables (“EmployeeCount”, “Over18”, “StandardHours”). Target variable is “Attrition”. Key variables are Environmental Satisfaction, Salary/Hourly Rate, Job Role/Satisfaction, Performance Ratio, Marital Status and so on. We also use descriptive statistics to get general information for each variable as shown in Figure 1. Furthermore, we obtain the distribution of each variable (Appendix1).

	count	mean	std	min	25%	50%	75%	max
Age	1470.0	36.92	9.14	18.0	30.00	36.0	43.00	60.0
DailyRate	1470.0	802.49	403.51	102.0	465.00	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.19	8.11	1.0	2.00	7.0	14.00	29.0
Education	1470.0	2.91	1.02	1.0	2.00	3.0	4.00	5.0
EmployeeCount	1470.0	1.00	0.00	1.0	1.00	1.0	1.00	1.0
EmployeeNumber	1470.0	1024.87	602.02	1.0	491.25	1020.5	1555.75	2068.0
EnvironmentSatisfaction	1470.0	2.72	1.09	1.0	2.00	3.0	4.00	4.0
HourlyRate	1470.0	65.89	20.33	30.0	48.00	66.0	83.75	100.0
JobInvolvement	1470.0	2.73	0.71	1.0	2.00	3.0	3.00	4.0
JobLevel	1470.0	2.06	1.11	1.0	1.00	2.0	3.00	5.0
JobSatisfaction	1470.0	2.73	1.10	1.0	2.00	3.0	4.00	4.0
MonthlyIncome	1470.0	6502.93	4707.96	1009.0	2911.00	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.10	7117.79	2094.0	8047.00	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.69	2.50	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.21	3.66	11.0	12.00	14.0	18.00	25.0
PerformanceRating	1470.0	3.15	0.36	3.0	3.00	3.0	3.00	4.0
RelationshipSatisfaction	1470.0	2.71	1.08	1.0	2.00	3.0	4.00	4.0
StandardHours	1470.0	80.00	0.00	80.0	80.00	80.0	80.00	80.0
StockOptionLevel	1470.0	0.79	0.85	0.0	0.00	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.28	7.78	0.0	6.00	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.80	1.29	0.0	2.00	3.0	3.00	6.0
WorkLifeBalance	1470.0	2.76	0.71	1.0	2.00	3.0	3.00	4.0
YearsAtCompany	1470.0	7.01	6.13	0.0	3.00	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.23	3.62	0.0	2.00	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.19	3.22	0.0	0.00	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.12	3.57	0.0	2.00	3.0	7.00	17.0

Figure 1: Descriptive Statistics

2.3 Visulization

In Figure 2, 16.12% of employees will quit and 83.88% of them will not quit. Employees who want to quit are much smaller than those who want to keep working. To get more insights, we visualized the data in four parts, including Demographic Attributes, Rewards/Salary, Job Environment/Performance and Overtime Work.

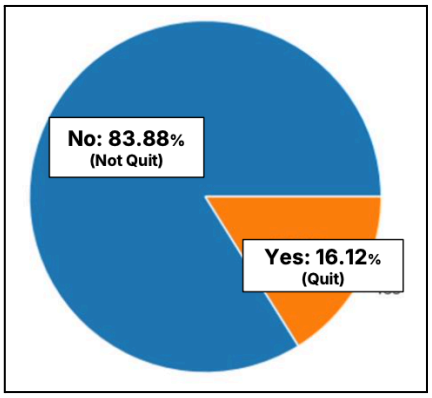

Figure 2: Attrition

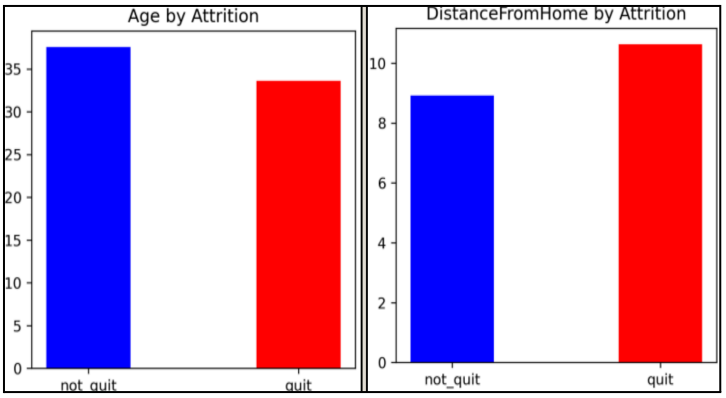

Figure 3: Demographic Attributes

In Figure 3, compared to those who "not quit", employees who "quit" are younger on average. The average age for employees who "quit" is around 30 years, while the average age for those who "not quit" is around 35 years, which shows that younger people are more likely to quit for their job. Besides, employees who "quit" are inclined to live farther away from their workplace compared to those who "not quit". Employees who quit live an average of 9 miles from work, whereas those who stay live about 7 miles away, which demonstrates that people living farther away from their workplace tend to quit for the job.

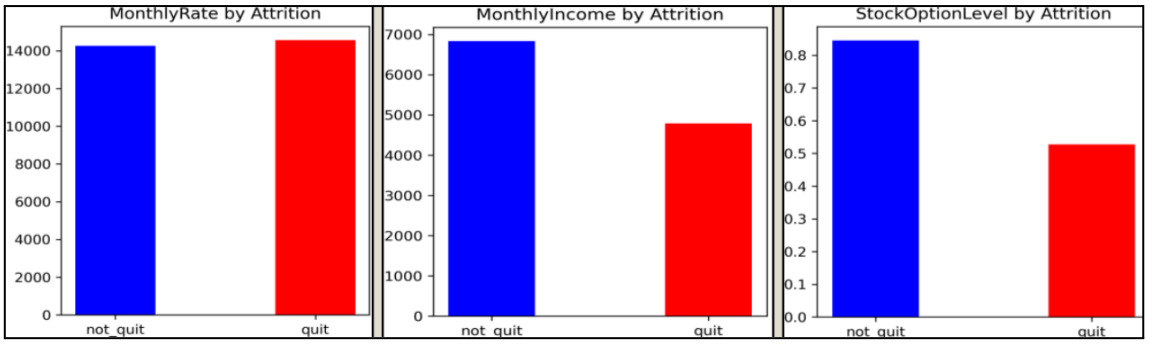

Figure 4: Rewards/Salary

In Figure 4, there is not a significant difference between MonthlyRate and Attrition. However, for Monthly Income and Stock Option Level, employees with lower incomes (around $5,000) and lower stock option levels (around 0.5) are more likely to quit. In contrast, employees who do not quit tend to have a monthly income close to $7,000 and a stock option level above 0.8.

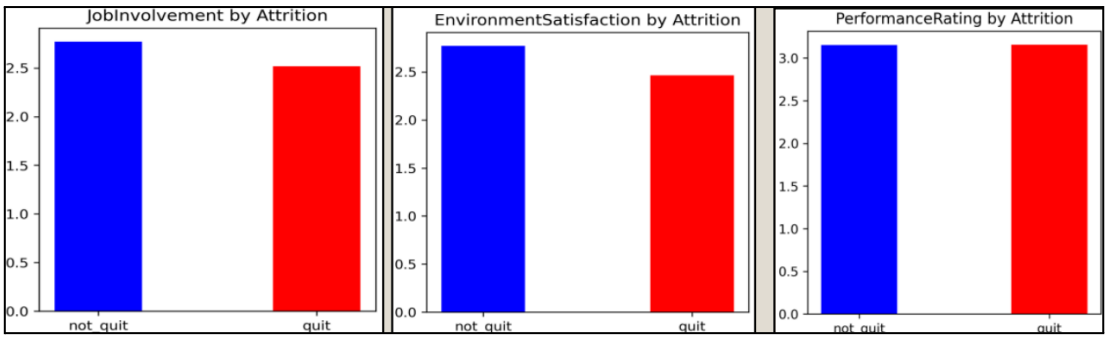

Figure 5: Job Environment/Performance

In Figure 5, employees who are not satisfied with their work environment and have lower levels of job involvement are more likely to quit. Additionally, there is no significant difference between performance rating and attrition.

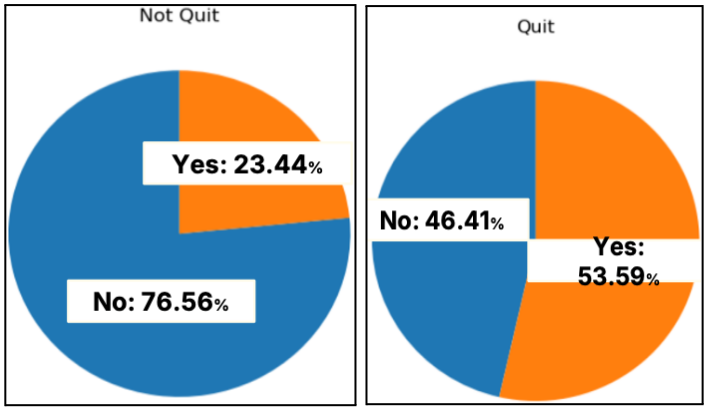

Figure 6: Overtime Work

In Figure 6, the majority of employees (76.56%) who "not quit" are inclined to arrive at work and leave on time, with only about 23.44% working overtime. By contrast, the majority of employees (53.59%) who "quit" tend to work overtime, suggesting that reducing overtime may improve employee retention.

3. Model Building

3.1 Initial Modeling

Before adjusting any of the data and parameters of the modeling, the proportions of the y variable are important to show random guessing accuracy as well as to benchmark how the models are doing initially. The Attrition variable is a two-class target variable that consists of Class 0 (No): 83.88% and Class 1 (Yes): 16.12%. In addition, most of the models did very well in terms of accuracy in the initial running, as can be seen below in Figure 7.

Model: Random Forest

	precision	recall	f1-score	support
0	0.99	0.87	0.93	432
1	0.08	0.56	0.14	9
accuracy			0.86	441
macro avg	0.54	0.71	0.53	441
weighted avg	0.97	0.86	0.91	441

Model: Logistic Regression

	precision	recall	f1-score	support
0	1.00	0.86	0.93	441
1	0.00	0.00	0.00	0
accuracy			0.86	441
macro avg	0.50	0.43	0.46	441
weighted avg	1.00	0.86	0.93	441

Model: K Nearest Neighbors

	precision	recall	f1-score	support
0	0.94	0.87	0.90	412
1	0.11	0.24	0.16	29
accuracy			0.83	441
macro avg	0.53	0.56	0.53	441
weighted avg	0.89	0.83	0.85	441

Figure 7: Initial Model performance

There was one preprocessing step done here that does not affect the scoring of the models, and that is changing the object type variables to integer types so that the models can actually process the data. The accuracies of the three models we chose were all at 0.83 or above. This may seem like a good thing, but as we take a closer look at the classification reports, Class One seems to perform very poorly relative to Class Zero. Immediately we can assume that this is due to an imbalance of proportion for the target variable. This means that we have to start several preprocessing methods and model tuning to help bring up the performance of the models when predicting Class One. To measure our benchmarks, we specifically want to target the performance of Class One's f1-score, which combines the precision and recall performance.

3.2 Preprocessing and Optimizing the Models

OneHotEncoding and Standardization: The first step of enhancing the model is the preprocessing step. The only part was to structure the data since there were no null values or duplicates to be handled. Structuring is done after the train test split to avoid data leaks. It's necessary to use the One Hot Encoding method to create dummy variables for categorical variables that aren't binary, mainly because each value in the variable does not have more weight than the other values. After creating the dummy variables, scaling the data to have appropriate data for modeling is next. A fit and transform would be applied to the X train, and a transform would be applied to the X_test. Since most of the data was normally distributed, the standard scaler was the perfect option for standardizing the data. Here is where our benchmark happens.

Classification report for RandomForest:

	precision	recall	f1-score	support
No	0.87	0.99	0.93	380
Yes	0.67	0.10	0.17	61
accuracy			0.87	441
macro avg	0.77	0.55	0.55	441
weighted avg	0.84	0.87	0.82	441

Classification report for LogisticRegression:

	precision	recall	f1-score	support
No	0.90	0.95	0.93	380
Yes	0.55	0.36	0.44	61
accuracy			0.87	441
macro avg	0.73	0.66	0.68	441
weighted avg	0.85	0.87	0.86	441

Classification report for KNN:

	precision	recall	f1-score	support
No	0.88	0.97	0.92	380
Yes	0.50	0.16	0.25	61
accuracy			0.86	441
macro avg	0.69	0.57	0.59	441
weighted avg	0.83	0.86	0.83	441

Figure 8: Model Performance after scaling the data

After scaling the data, there is a significant increase in the f1-score of Class One for our models. Random Forest increases the f1-score by 0.03, Logistic Regression increases by 0.44, and KNN increases by 0.09. One positive note here is that most of the other performances, like overall accuracy and f1-score of Class Zero, were not compensated for the Class One performance increase. However, there are still more steps that we can take to improve the models further.

SMOTE: Since there is a heavy imbalance in the dataset, which was mentioned earlier, the SMOTE (Synthetic Minority Oversampling Technique) is the method to use to oversample the minority class of the target variable. The sampling strategy was set to auto, and then the scaled data was resampled using SMOTE. Again, we benchmark the performances.

Classification report for RandomForest:

	precision	recall	f1-score	support
No	0.88	0.98	0.93	380
Yes	0.58	0.18	0.28	61
accuracy			0.87	441
macro avg	0.73	0.58	0.60	441
weighted avg	0.84	0.87	0.84	441

Classification report for LogisticRegression:

	precision	recall	f1-score	support
No	0.93	0.78	0.85	380
Yes	0.31	0.61	0.41	61
accuracy			0.76	441
macro avg	0.62	0.69	0.63	441
weighted avg	0.84	0.76	0.78	441

Classification report for KNN:

	precision	recall	f1-score	support
No	0.91	0.61	0.73	380
Yes	0.21	0.64	0.31	61
accuracy			0.61	441
macro avg	0.56	0.62	0.52	441
weighted avg	0.81	0.61	0.67	441

Figure 9: Model Performance after SMOTE

Here is where we start seeing a drop off in other metrics to compensate for the increase of metrics in Class One. Random Forest increases Class One f1-score by 0.11, Logistic Regression decreases by 0.03, and K Nearest Neighbors increases by 0.06. Only the Random Forest keeps the accuracy up, while the other models have some dropoff. However, if we set a threshold of 0.75 for f1-score, then all of the models aren't underperforming too much for Class Zero.

Principal Component Analysis: PCA is a common method to reduce the dimensions of the features in the models while still retaining the variance. The number of components has been set to 33, which is 2 components less than the number of features. This technique is useful because it reduces the complexity of the dataset, and it does increase most of the metrics in the classification reports of our models.

Classification report for RandomForest:

	precision	recall	f1-score	support
No	0.88	0.97	0.92	380
Yes	0.45	0.15	0.22	61
accuracy			0.86	441
macro avg	0.66	0.56	0.57	441
weighted avg	0.82	0.86	0.82	441

Classification report for LogisticRegression:

	precision	recall	f1-score	support
No	0.94	0.78	0.85	380
Yes	0.33	0.69	0.45	61
accuracy			0.77	441
macro avg	0.64	0.73	0.65	441
weighted avg	0.86	0.77	0.80	441

Classification report for KNN:

	precision	recall	f1-score	support
No	0.90	0.61	0.73	380
Yes	0.20	0.59	0.30	61
accuracy			0.61	441
macro avg	0.55	0.60	0.51	441
weighted avg	0.81	0.61	0.67	441

Figure 10: Model Performance after PCA

Unfortunately, we can see a dropoff in the Random Forest in the entire classification report after PCA. Because of this, we will not be using the dimension-reduced variables when moving on. And same for K Nearest Neighbors; there was a small drop off as well, so we will not be using dimension-reduced variables when moving on. However, in Logistic Regression, there is a 0.04 increase in f-1 score for Class One and an increase in overall accuracy of 0.01 to 0.77. These results determine that we will only use PCA for our Logistic Regression model going forward into our final step, using the GridSearch with 5 Fold Cross Validation to find the best parameters for our models.

3.3 Random Forest (Final Model)

GridSearch: The following are the best parameters for the random forest model, derived from the GridSearch: {'bootstrap': False, 'criterion': 'gini', 'max_depth': 30, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200, 'oob_score': False}. Using these parameters, without PCA, the following is the improved classification report:

	precision	recall	f1-score	support
No	0.98	0.88	0.93	423
Yes	0.18	0.61	0.28	18
accuracy			0.87	441
macro avg	0.58	0.75	0.60	441
weighted avg	0.95	0.87	0.90	441

Figure 11: Model Performance of Random Forest with Grid Search method

The accuracy rises back to 0.87, and the f1-score for Class Zero is retained. However, it seems that the recall numbers and precision numbers are swapped after the Grid Search. This could mean that there is no more improvement available for the model. The Random Forest Model before Grid Search can be used if the Class One precision needs to be the highest, and the current Model here should be used if the recall needs to be maximized. In our case, we recommend increasing recall (Grid Search Model) because this would allow for maximum Recall. It is fine for the company to predict false positives for an employee that wants to resign if the company wants to maximize the employee retention rate.

3.4 Logistic regression (Final Model)

GridSearch: The following are the best parameters for the logistic regression model, derived from the GridSearch: {'C':1, 'max_iter':100, 'penalty':'l1','solver':'liblinear'}. Using these parameters, with PCA, the following is the improved classification report:

	precision	recall	f1-score	support
No	0.78	0.94	0.85	315
Yes	0.69	0.33	0.45	126
accuracy			0.77	441
macro avg	0.73	0.64	0.65	441
weighted avg	0.75	0.77	0.74	441

Figure 12: Model Performance of Logistic Regression with Grid Search method

The accuracy stays the same compared to before the Grid Search, and the near identical phenomenon to the Grid Search Random Forest happens here, where the precision and recall numbers are swapped after the Grid Search. However, in this case, it is quite the opposite. The model before Grid Search has the higher precision, so it is recommended to use the previous PCA/Default Logistic Regression Model for company use in employee retention.

3.5 K Nearest Neighbors (Final Model)

GridSearch: The following are the best parameters for the K Nearest Neighbors model, derived from the GridSearch: {'metric':'minkowski', 'n_neighbors': 3, 'weights': 'distance'}. Using these parameters, without PCA, the following is the improved classification report:

	precision	recall	f1-score	support
No	0.62	0.90	0.74	263
Yes	0.57	0.20	0.29	178
accuracy			0.62	441
macro avg	0.60	0.55	0.52	441
weighted avg	0.60	0.62	0.56	441

Figure 13: Model Performance of K Nearest Neighbors with Grid Search method

The accuracy increases by 0.01 to 0.62 for K Nearest Neighbors after Grid Search. Again, the recall and precision are swapped after the Grid Search, so it is a preference option for which model to choose. The model without PCA/Default KNN is recommended for companies trying to achieve employee retention.

3.6 Comparison (Expected values based on confusion matrix)

An expected value framework can be derived from the confusion matrix of the predicted outcomes of the models. We took the example of a software company to create this framework. We defined the cost of a single employee turnover to be $125,000 (Yearly), based on factors like productivity, training, onboarding, and recruitment costs in looking for new employees. The cost of employee retention comes from the most important features coming from the best-performing model: the Random Forest Classifier.

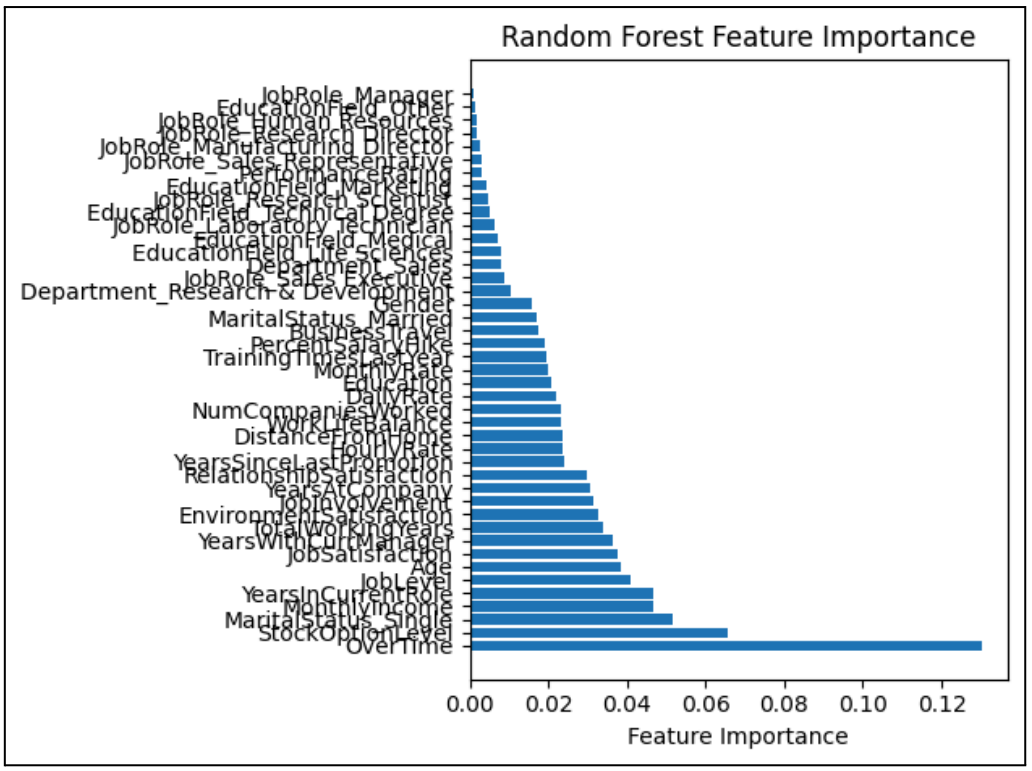

Figure 14: Random Forest Feature Importance

Factors like bonuses, overtime, stock options, and a better work environment account for around $10,000 for each employee (Yearly). By taking the difference of these two factors, we get the benefit of retaining an employee who wants to resign: $115,000. These values can be multiplied by the predicted and true labels from the confusion matrices of the models.

```
rf_tn, rf_fp, rf_fn, rf_tp = rfcm.ravel()
total = rf_tn+rf_fp+rf_fn+rf_tp
115000*rf_tp/total-10000*rf_fp/total-125000*rf_fn/total
```

```
-249.43310657596385
```

```
lr_tn, lr_fp, lr_fn, lr_tp = lrcm.ravel()
115000*lr_tp/total-10000*lr_fp/total-125000*lr_fn/total
```

```
-13287.98185941043
```

```
knn_tn, knn_fp, knn_fn, knn_tp = knncm.ravel()
115000*knn_tp/total-10000*knn_fp/total-125000*knn_fn/total
```

```
-31995.46485260771
```

(1 Random Forest, 2 Logistic Regression, and 3 K Nearest Neighbors)
Figure 15: Expected Value for Each Model

As shown in Figure 15, for the Random Forest Classifier, there is an expected loss of $249.43 for each employee retention. For Logistic Regression Classifier, there is an expected loss of $13,287.98 for each employee retention. For K Nearest Neighbors, there is an expected loss of $31,995.46 for each employee retention. These amounts are heavily dependent on the accuracy and other metrics from the classification reports, with K Nearest Neighbors performing the worst out of the three models.

3.7 Limitations

1. Imbalanced Dataset: The dataset has a significant class imbalance (16.12% attrition vs. 83.88% retention), which impacts the performance of models, especially in predicting the minority class (attrition, yes). Despite using SMOTE, some models still struggle to achieve a balance between precision and recall for the minority class.
2. Limited Class One Predictive Performance: One key limitation is the inability to increase the f1-score for Class One (Yes for Attrition) to a level that would be meaningful for real-world predictions. A larger sample size with more instances of Class One could have improved model performance, as the minority class lacked sufficient representation for effective learning.
3. Generalization Issues: The dataset is sourced from a single organization (IBM) and may not generalize well to other companies with different cultures, policies, or employee dynamics. The insights may be specific to this dataset and not applicable across diverse industries.
4. Limited Feature Scope: While the dataset includes 35 features, it may lack certain key factors influencing attrition, such as external market conditions, economic trends, or detailed performance metrics like peer reviews, which could enhance the predictive power of the models.
5. Overfitting Risks: With extensive hyperparameter tuning (e.g., GridSearch), there is a risk of overfitting the models to this specific dataset, which could reduce their performance on unseen data.
6. Over Reliance on Quantitative Data: The analysis relies heavily on quantitative data and does not incorporate qualitative insights, such as employee feedback or exit interviews, which could provide a more holistic understanding of attrition causes.

4. Takeaways

In the process of visualizing the data, we gained deeper insights into the key factors driving employee attrition, which provided robust support for our recommendations. Demographic characteristics emerged as significant indicators; younger employees demonstrated a significantly higher attrition rate compared to older employees. Employees around the age of 30 were more likely to leave, while those aged 35 and above showed a higher tendency to stay. Similarly, the distance between an employee's residence and the workplace proved to be a crucial factor. Employees living approximately 9 miles away from the workplace had a higher likelihood of leaving than those living within 7 miles.

Regarding compensation and rewards, while monthly income levels did not show a strong direct correlation with attrition, employees with lower income and fewer stock options exhibited a higher propensity to leave. For instance, employees earning around $5,000 per month with limited stock options had a higher attrition rate compared to those earning $7,000 per month and receiving more stock options. This highlights the importance of offering competitive compensation packages and long-term incentives to enhance retention.

Work environment and engagement also played significant roles. Dissatisfaction with the work environment and lower levels of engagement were strongly associated with higher attrition rates. Interestingly, performance evaluation scores showed no significant correlation

with attrition, suggesting that subjective perceptions and environmental factors were more influential in an employee's decision to leave.

Additionally, excessive overtime emerged as an issue warranting attention. Employees who frequently worked overtime were more likely to leave compared to those with regular work hours. This reinforces the critical importance of promoting work-life balance. Reducing overtime and offering flexible work arrangements could directly improve employee satisfaction and reduce attrition.

SMOTE helped the most in improving the model performances because of its ability to oversample the minority class, creating an even distribution of the target variable. Both the original model's benchmarks and the target variable's proportion, which means that there was some success in the preprocessing and optimizing of the models.

The Logistic Regression model had the most improvement in the f1-score, from 0.00 initially to 0.45 after preprocessing and model building. However, this is still under an ideal threshold, where we believe there needs to be more Class 2 samples in the data. Precision should be maximized in this business problem, which makes the Random Forest Classifier the best model for this problem. In addition, the net cost for employee retention is the lowest for this classifier as well. The K Nearest Neighbors model is not recommended, as it will cost a company the most for employee retention with its prediction.

In summary, our analysis highlights key drivers of employee attrition, including age, commute distance, compensation, work environment, and overtime. Among the models tested, the Random Forest classifier proved to be the most effective, offering the best balance of precision and cost efficiency for retention efforts. These findings underscore the importance of targeted strategies to improve work-life balance, enhance compensation, and address engagement to reduce attrition effectively.

5. Conclusions

The Random Forest classifier stood out with an accuracy of 0.87, making it the most effective model for predicting employee attrition. It also unveiled crucial influencing factors, such as compensation structure, stock option levels, work-life balance, and the distance between an employee's residence and workplace. These findings enabled us to recommend actionable measures, including optimizing salary policies, enhancing the work environment, reducing overtime, and supporting younger employees and those living farther away from the office, including enhancing remote work.

However, the limitations of this study should not be overlooked. The imbalance in the target variable within the dataset impacted the model's ability to predict attrition accurately. Additionally, the inability to quantify psychological and personal factors limited the comprehensiveness of our analysis. Future research could address these gaps by expanding the sample size and integrating qualitative data to further refine the model's applicability.

In summary, this study demonstrates how machine learning techniques can empower organizations to predict and reduce employee attrition. By proactively addressing attrition

risks, companies can lower costs, stabilize teams, and create more attractive work environments, thereby achieving long-term competitive advantages. We believe these data-driven insights will provide essential support for strategic decision-making and drive sustained organizational success.

Appendix

- Distribution of each variable