FINAL PROJECT REPORT

ISM 6353

Professor Andy Chen

Yayuan Zhang, Peiling Zou

Contents

Abstract

Why employees leave has always been a difficult problem for companies. With regard to understanding how to reduce the employee turnover rate, it is necessary to start by studying why employees leave. For each enterprise, controlling the employee turnover rate can effectively control the enterprise's labor cost.

Employee turnover rate is an important indicator of human resource management. The dataset for this analysis comes from fictitious employee turnover data created by IBM data scientists. In this article, the following will be carried out through the analysis of HR employee turnover data.

- A quick visualization and exploratory analysis of some important variables, especially those related to essential information, monthly income, promotion, satisfaction, performance, and work-life balance.
- Analyze the factors that contribute to employee turnover and explore the degree of influence of each variable.
- Construct models through effective algorithms for predicting whether an employee will resign or not.

With respect to this post, our goal is to use the process and results of the analysis to help reduce employee turnover when modeling predictive analysis using data sets. With this, it can help organizations understand the attraction and satisfaction of employees to the organization and allow management to clarify what factors influence employee turnover. We

conclude the article with some recommendations based on what the dataset tells us about the story and the predictive analytics model.

Business insight

1.1 Business Problems and Motivations

This article will examine and make predictions about employee turnover. Why is employee turnover so important? One of the biggest problems for many businesses is employee turnover. Companies that experience it may lose employee productivity, have to acquire new workers, have reduced morale, lose out on sales possibilities, and incur additional costs. The time and money spent on employee training is a significant expense for businesses.

At the same time, a senior employee's leaving leaves a corporate hole that harms the organization's reputation in addition to its stability. If they had initially apprehended the employee, they may have averted this predicament. Companies need to understand the real drivers behind turnover rates based on employee data and characteristics and provide the actionable/feasible insights they need to better understand their employees. In addition, the results of the model analysis will be able to help companies quickly identify potential issues and what dimensions and aspects can be used to improve employee retention.

1.2 Stakeholders

Organization: C-suite management, executives, managers, HR.

Companies need to understand what causes employee turnover, whether it's dissatisfaction with the work environment, wages, or management issues at managers that lead to turnover. Based on the model analysis and data visualization, it helps companies and

executives quickly identify potential problems and resignation trends, improve management issues with corresponding deficiencies and increase employee loyalty.

1.3 Source

Data on employee features was obtained from a fictional data set created by IBM data scientists and published on Kaggle. The primary application is to reveal the factors that contribute to employee turnover and to explore the impact that multiple factors have on employee performance.

1.4 Data

The dataset contains 1471 data points and 35 features. Each data point represents the employee's basics information and work-related information, such as age, gender, department, education, hourly rate, distance from home, job involvement, job level, marital status, monthly income, relationship satisfaction, environment satisfaction, performance rating, total working years, years at the company, work-life balance, training times last year, years since last promotion.

2. Data Analysis

2.1 Workflow Analysis process

Throughout the workflow analysis process, firstly we import libraries and datasets to read the data, such as checking for missing values, checking data types, performing necessary generalizations, and data cleaning. After we understood what the data looked like, we explored the impact of multiple factors on employee turnover and performed EDA (Exploratory data analysis) and predictive modeling to uncover potential issues and where turnover could be reduced.

In the predictive modeling process, we split the observation class domain into 3 datasets in the submission and tested 4 different models. When it comes to our data, using domain data types rather than base data types guarantees that we preserve consistency across an organization while also enabling the reuse of standard data type definitions for increased team productivity.

2.2 Data Description

Our target feature is each column, and the predictor variable is Attrition. In the predictive model, we used some important variables, especially those related to essential information, monthly income, promotion, satisfaction, performance, and work-life balance.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 36 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   index                     1470 non-null   int64
 1   Age                       1470 non-null   int64
 2   Attrition                 1470 non-null   object
 3   BusinessTravel            1470 non-null   object
 4   DailyRate                 1470 non-null   int64
 5   Department                1470 non-null   object
 6   DistanceFromHome          1470 non-null   int64
 7   Education                 1470 non-null   int64
 8   EducationField            1470 non-null   object
 9   EmployeeCount             1470 non-null   int64
 10  EmployeeNumber            1470 non-null   int64
 11  EnvironmentSatisfaction   1470 non-null   int64
 12  Gender                    1470 non-null   object
 13  HourlyRate                1470 non-null   int64
 14  JobInvolvement            1470 non-null   int64
 15  JobLevel                  1470 non-null   int64
 16  JobRole                   1470 non-null   object
 17  JobSatisfaction           1470 non-null   int64
 18  MaritalStatus             1470 non-null   object
 19  MonthlyIncome             1470 non-null   int64
 20  MonthlyRate               1470 non-null   int64
 21  NumCompaniesWorked        1470 non-null   int64
 22  Over18                    1470 non-null   object
 23  OverTime                  1470 non-null   object
 24  PercentSalaryHike         1470 non-null   int64
 25  PerformanceRating         1470 non-null   int64
 26  RelationshipSatisfaction  1470 non-null   int64
 27  StandardHours             1470 non-null   int64
 28  StockOptionLevel          1470 non-null   int64
 29  TotalWorkingYears         1470 non-null   int64
 30  TrainingTimesLastYear     1470 non-null   int64
 31  WorkLifeBalance           1470 non-null   int64
 32  YearsAtCompany            1470 non-null   int64
 33  YearsInCurrentRole        1470 non-null   int64
 34  YearsSinceLastPromotion   1470 non-null   int64
 35  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(27), object(9)
memory usage: 413.6+ KB
```

After examining the data information, we learned that there are no null values in the entire database, there are 9 OBJECTs out of 36 features, and the rest are all INTs. The lack of nulls is beneficial to us since eliminating nulls from the dataset is an essential step in data contention before beginning EDA and constructing predictive models. Because the performance and accuracy of any machine learning algorithm might be adversely affected by these null values. As a result, it is crucial to exclude the null values from the dataset before using any machine learning method on it. The other good news is the dataset has no data quality issues.

	count	unique	top	freq
Attrition	1470	2	No	1233
BusinessTravel	1470	3	Travel_Rarely	1043
Department	1470	3	Research & Development	961
EducationField	1470	6	Life Sciences	606
Gender	1470	2	Male	882
JobRole	1470	9	Sales Executive	326
MaritalStatus	1470	3	Married	673
Over18	1470	1	Y	1470
OverTime	1470	2	No	1054

	index	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000
mean	734.500000	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.891156
std	424.496761	9.135373	403.509100	8.106864	1.024165	0.0	602.024335	1.093082	20.329428
min	0.000000	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000
25%	367.250000	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.000000
50%	734.500000	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.000000
75%	1101.750000	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000	4.000000	83.750000
max	1469.000000	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.000000

Afterwards, we quickly glanced at the summary statistics and the description of the data. The table revealed a variety of intriguing details and some of which may be the subject of the EDA session. The ratio of non-turnover to turnover is about 5:1. The average age of employees is about 36 years old. The average income of employees is about $ 6,500, with a median of US$ 4,919, which is more reflective of the salary level of the company. The longest period in this data set without a promotion is 18 years, and the longest period under the same management is 17 years. Employees in this data collection have a maximum of 40 years of experience. The oldest employee was 60 years old, and 29 miles from home was the greatest distance.

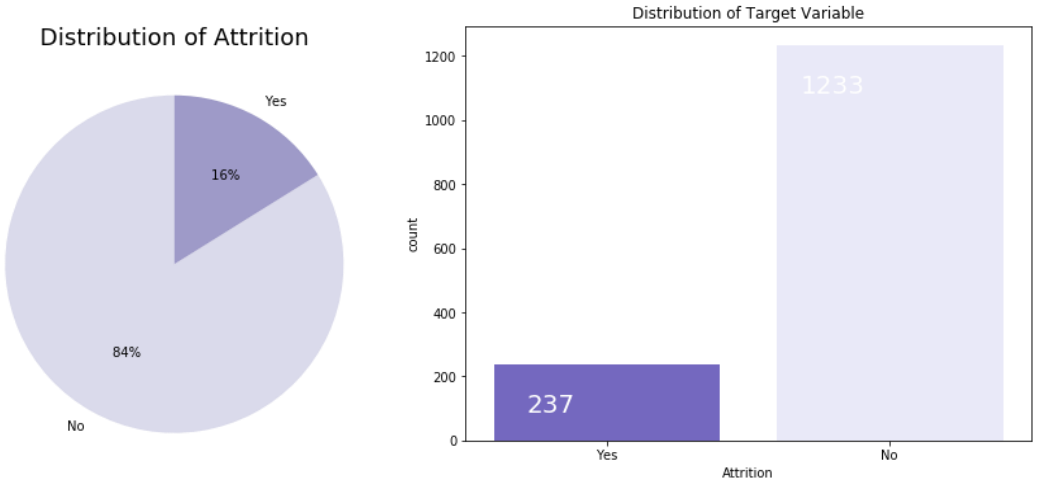
2.3 Data Preparation

We cleaned up the data before moving on to EDA and predictive modeling. In the process of data cleaning, incorrect, poorly formatted, or otherwise confusing data are sorted and corrected. Duplicate or irrelevant observations are removed, structural errors are fixed, string data types are changed to numeric types, incorrect data types are changed to correct types, unwanted outliers are filtered, missing data is handled, and the data's consistency is checked. Since we found in the previous step that the Attrition variable shows Yes and No,

which will not be good for our analysis later. Therefore, we need to convert the string data type of the Attrition variable to a numeric type.

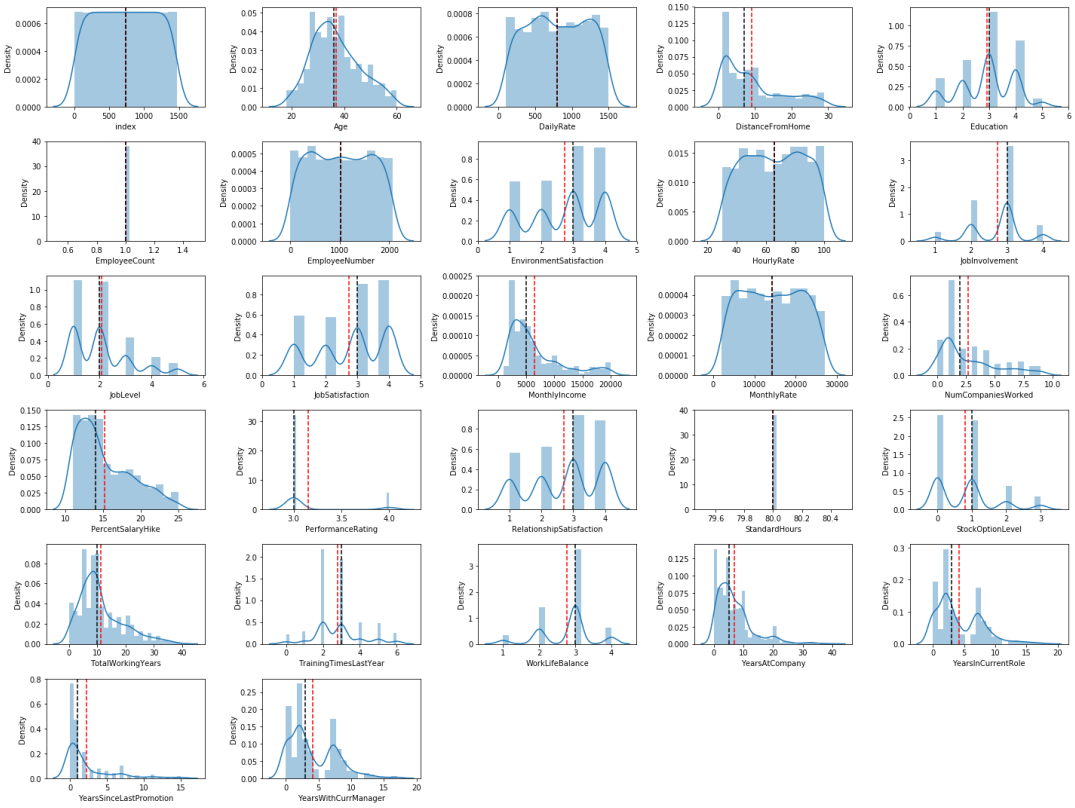
2.4 Explanatory Data Analysis (EDA)

2.4.1 Attrition variable



In this dataset, 84% of employees stay in the company and 16% of employees left the company, which is 237 out of 1470 employees. We noticed our target was unbalanced, which may cause a problem of instability in the dataset, as it clearly favors employees who choose to stay with the company.

2.4.2 Bar chart for density of each variable



We made a bar chart based on the density values of each variable. In this dataset, most employees are in their 30s. Most of the employees have 6-15 years of working experience. 63 employees worked for more than 28 years. Most employees stay with the company for 3-9 years, with a median of 5 years. 104 employees who have been with the company for more than 18 years. Most employees were promoted within 0-3 years, and 107 employees were not promoted for more than 7 years. Employee performance ratings are not the best with a median of about 3.0.

2.4.3 Heatmap Correlation

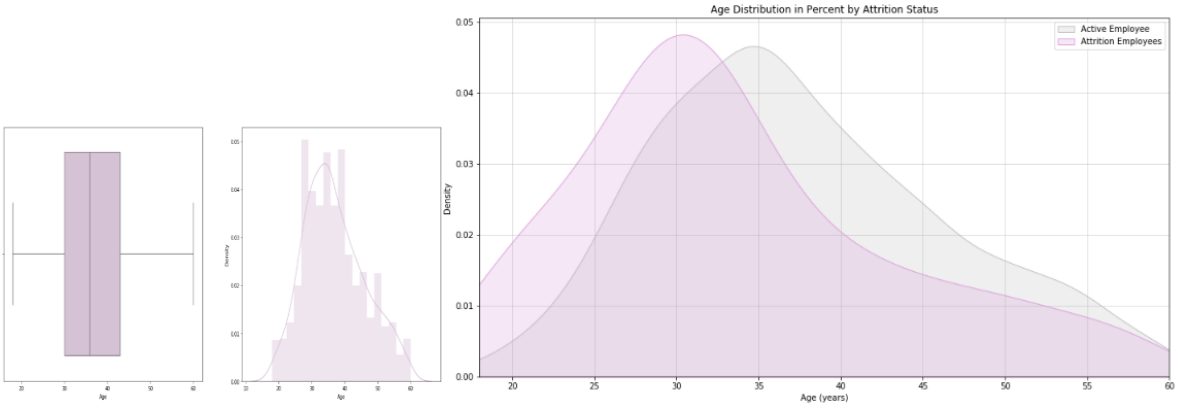
There is no strong correlation between the target column and any numeric columns. However, we can see other correlations such as -

- ★ It's clear that Senior employees have higher total working years.
- ★ Higher performance ratings lead to a higher percent salary hike.
- ★ The longer working years of Employees get more monthly income & have higher job levels

★ As the years go by, many employees remain in their current positions and under the same manager, which means they are not promoted.

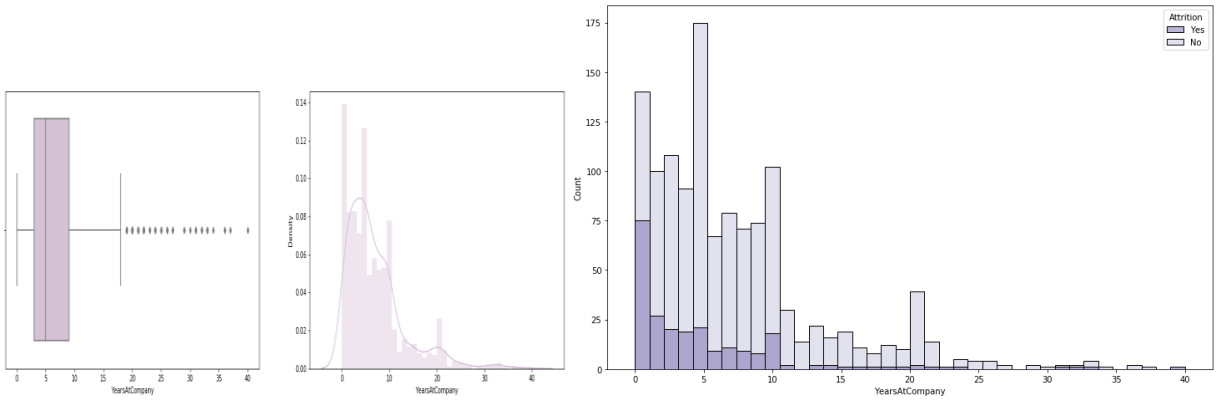
2.4.4 Explore the relationship between base information Gender, Age, Department, Job Level, Education and other variables and employee turnover

How does age contribute to attrition?



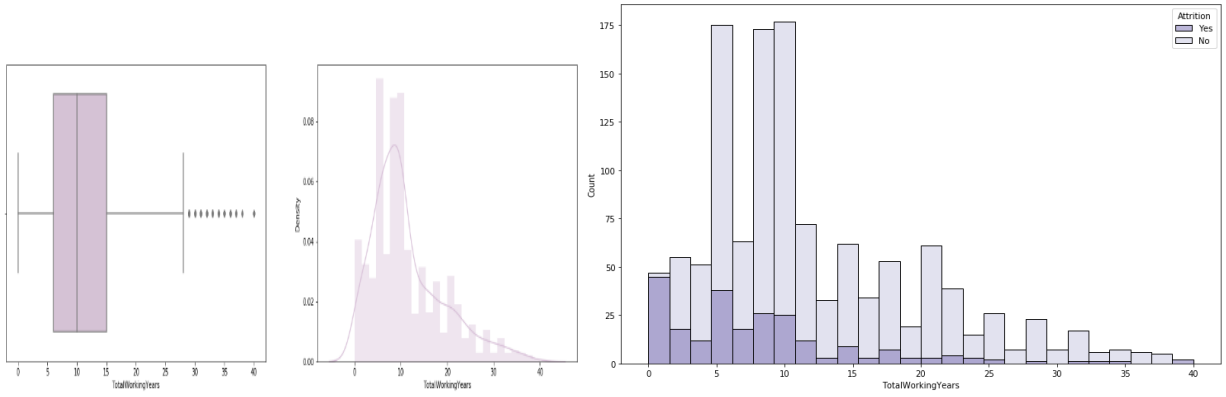
Through in-depth analysis of age. We learn that most employees are 35-year-olds. The median employee who leaves is 32 years old and the standard deviation is 9.7. The median current employee is 36 years old, and the standard deviation is 8.9. The data set has a high turnover rate at lower ages, mainly among employees younger than 30 years old.

How is Attrition affected by Year at Company?



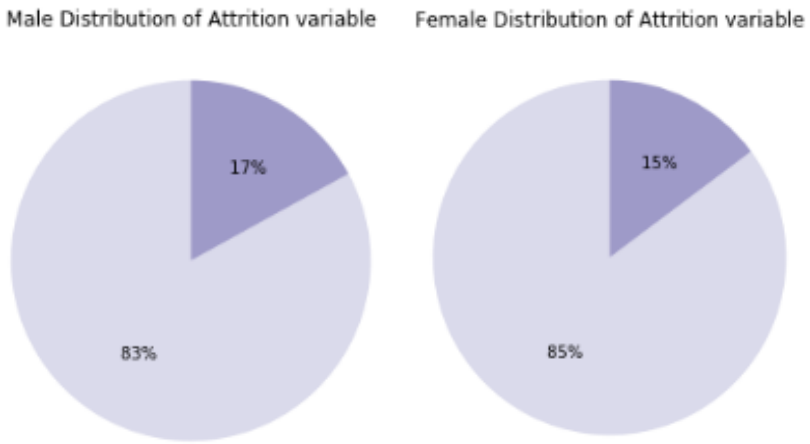
Through in-depth analysis and coding elaboration, we found that 104 employees worked in the company for more than 18 years. 580 employees worked in the company for less than 5 years. The turnover rate is high among employees who have been with the company for a short period of time, with a higher concentration of employees with less than 4 years of service.

Total Working Years



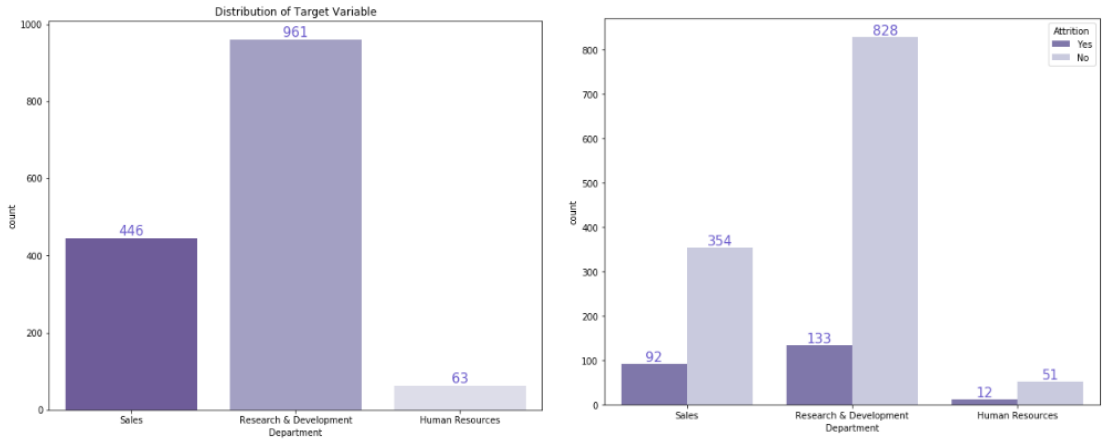
In this dataset, the majority of employees have 5-15 years of experience, with a median of 10 years. Through in-depth analysis and coding elaboration, we found that 63 employees have more than 28 years of experience. 484 employees have between 10 and 28 years of total experience. 831 employees have between 2 and 10 years of experience. 92 employees have just started working less than 2 years. The turnover rate is high for employees with a low length of service, concentrated in those with less than 7 years of service.

Gender



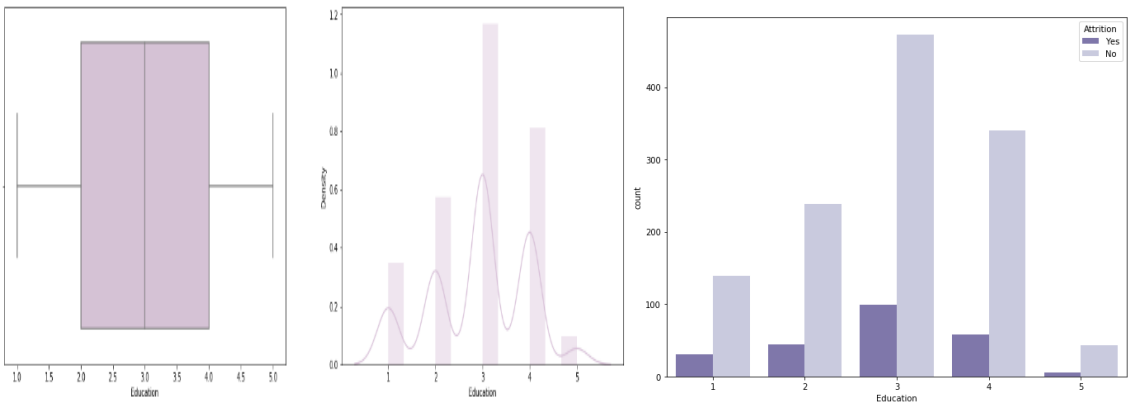
We can observe that there are more men than women in the company, and the turnover rate is slightly higher for men. Therefore, we do not consider gender as a major consideration due to there seems to be no difference in gender.

Department



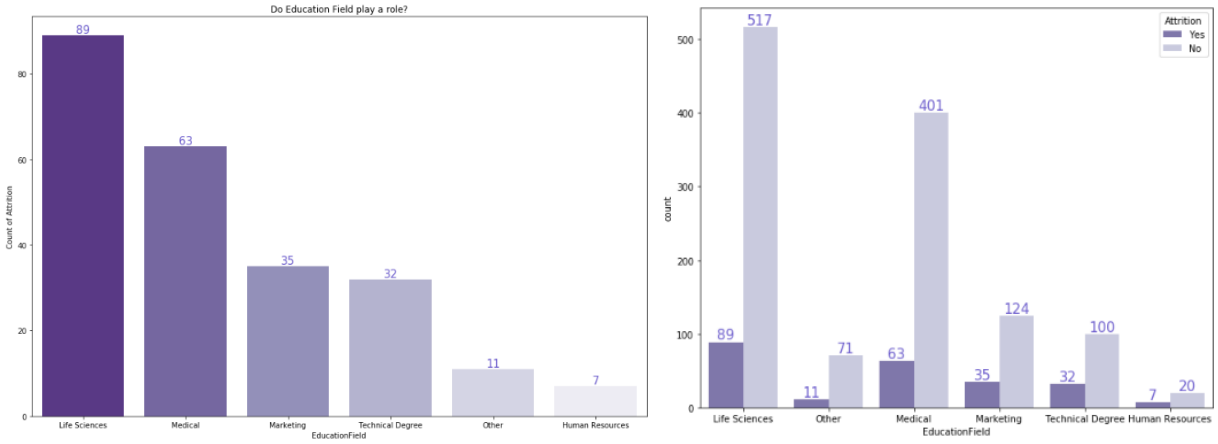
Most of the attrition came from the R & D department, with only Sales coming in second by a small margin. The lowest attrition rate was in human resources. However, we should consider how many people work in each department. Compared to Human resources and R&D departments. The sales department has the highest attrition when we look at the proportion.

Education



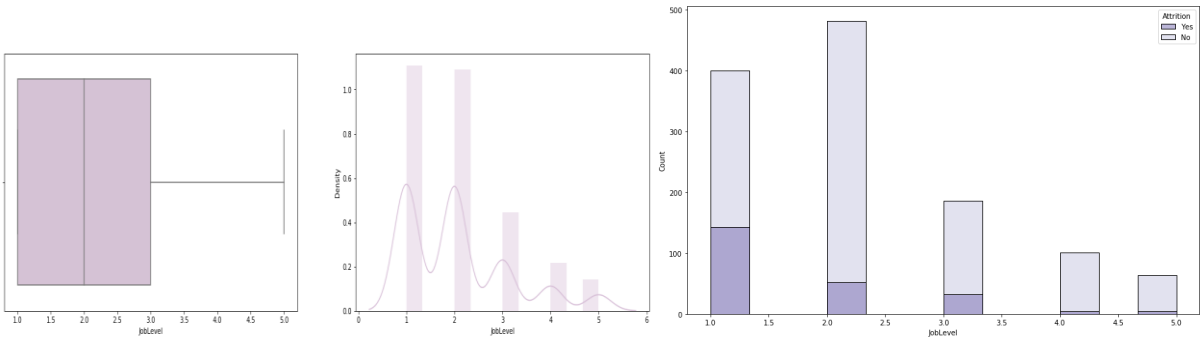
For the education variable, most employees are at levels 2 to 4, with a median of 3. Most workers are employed in level 3 schooling. Therefore, we do not consider Education as a major consideration due to there seems to be no difference in gender.

Education Field



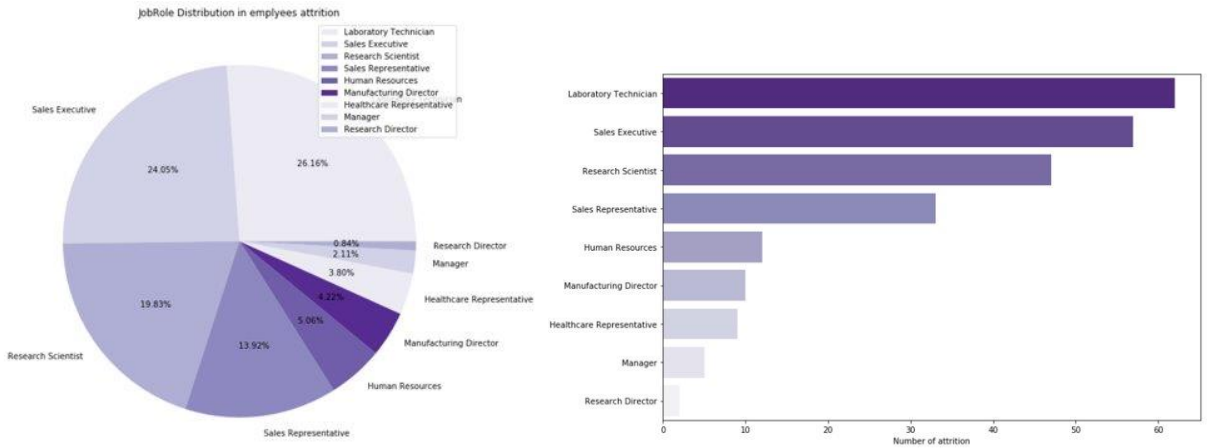
From the variables in education, we observed that employees in life sciences, medicine, and manufacturing leave the company. Through in-depth analysis, we found that there are many employees who study life sciences and medicine.

Job Level



Most employees have low job levels between 1-3, with a median of 2. Through in-depth analysis, we observed a high turnover rate of employees in low-ranking positions, mainly concentrated in positions with rank 1 level.

Job Role



Among job roles, most lab technicians, research scientists, sales executives and sales representatives choose to leave their jobs. Therefore, we considered delving into the salaries of each role to see if this was one of the main reasons.

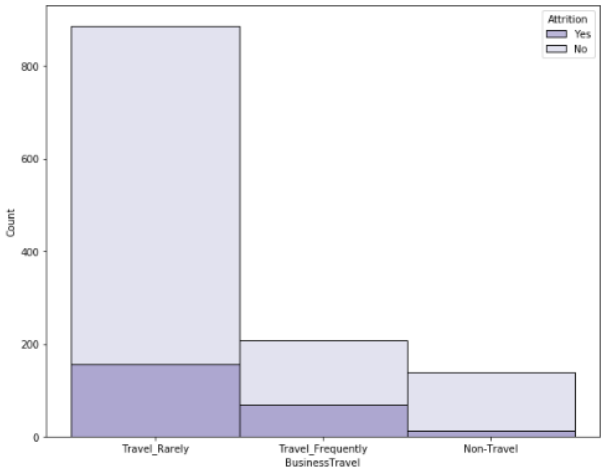
Summary

According to the age, total working time, and length of service of the lost workers, the likely explanation is that young people tend to try more and are somewhat uncertain about their future objectives, and the high turnover rate also suggests that it is difficult for such employees to create long-term identification with corporate values in the short term.

Concerning the job level, the turnover rate of each department corresponds to the frontline staff and may have some relationship with the nature of the business staff's work, how to minimize the turnover rate can focus on the Sales Department, in-depth study and excavation of probable causes.

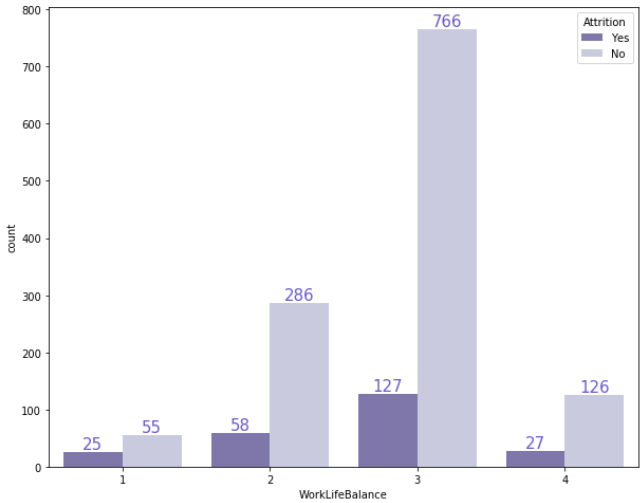
2.4.5. Explore the relationship between work-life balance-related variables and employee turnover

Business travel



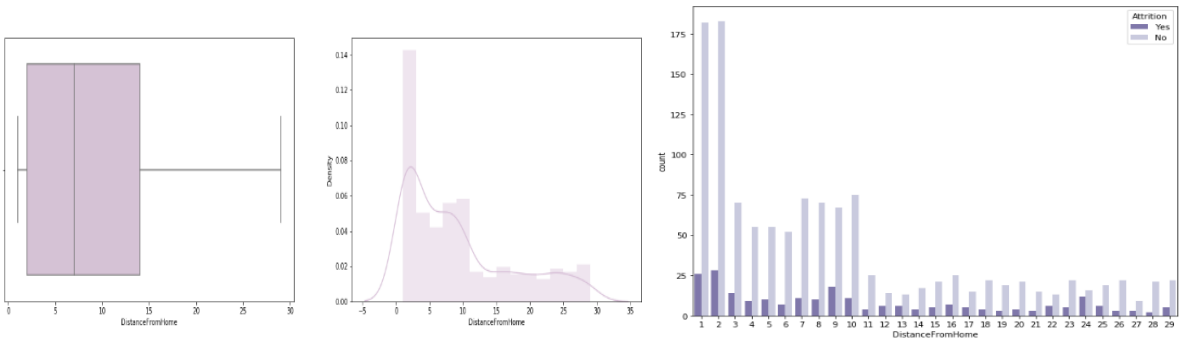
Most travelers practically never quit their employers. According to the plot, turnover is higher among employees who travel frequently in terms of percentage.

Work-life balance



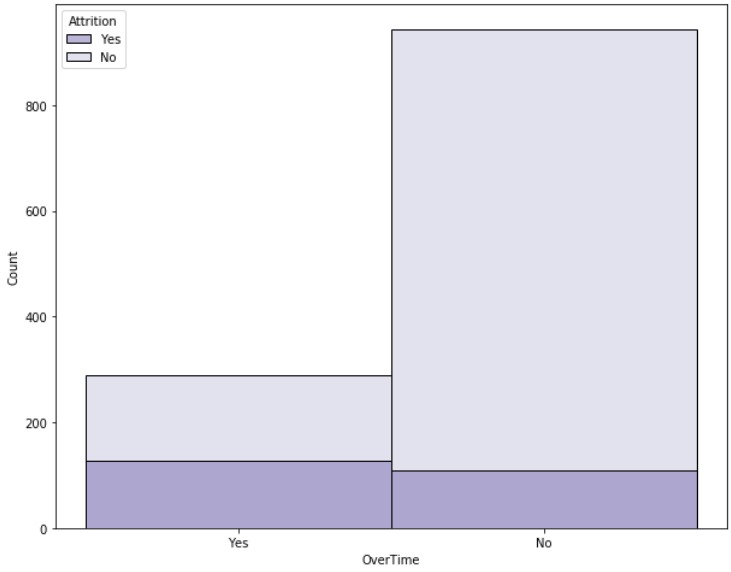
Employee turnover is high for those who believe that the work-life balance level is 1 in terms of percentage.

Distance from home



For distance from home, most employees are 2 miles to 14 miles distance from home, with 7 miles as the median. The turnover rate is higher among employees who are far from home in terms of percentage.

Over Time



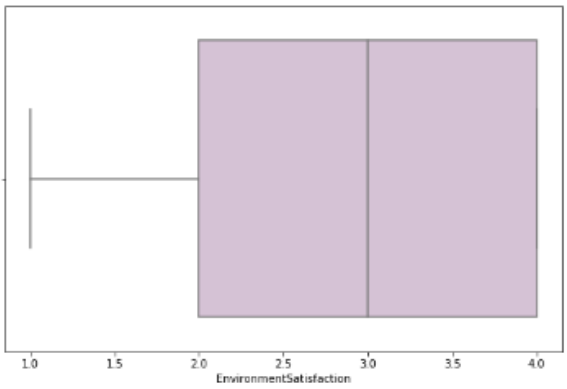
Employees who work overtime frequently have a very high turnover rate compared to those who do not work overtime in terms of percentage.

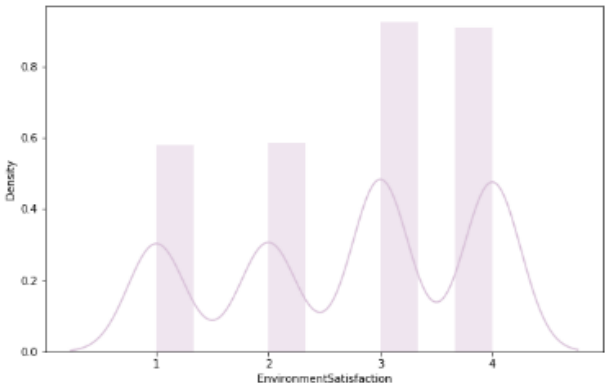
Summary

The greatest significant detriment to the quality of life is overtime, and the effects are much more obvious. Employees who travel more regularly are indeed easy to lose, and employees who go to work far away are also easy to lose, in general, the work-life balance of this sort of element has a more significant influence on employee turnover.

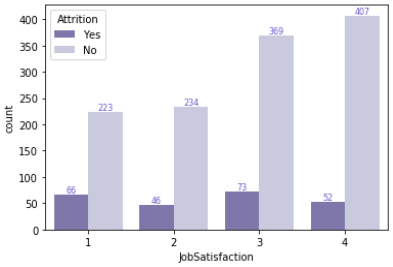
2.4.6 Exploring the relationship between satisfaction-related variables and employee turnover

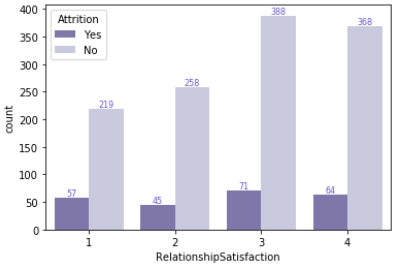
Environment Satisfaction, Relationship Satisfaction, Job Satisfaction

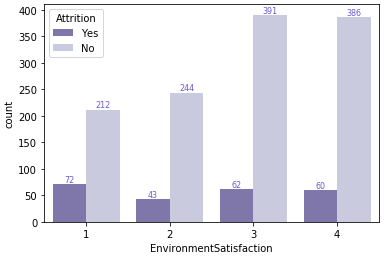




Based on the variable of environmental satisfaction, we observed that most employees felt satisfied with their work environment, with a median of 3.0. Through in-depth analysis and elaboration by EDA, we observed that the more satisfied employees are with their work environment, the less likely they are to leave.







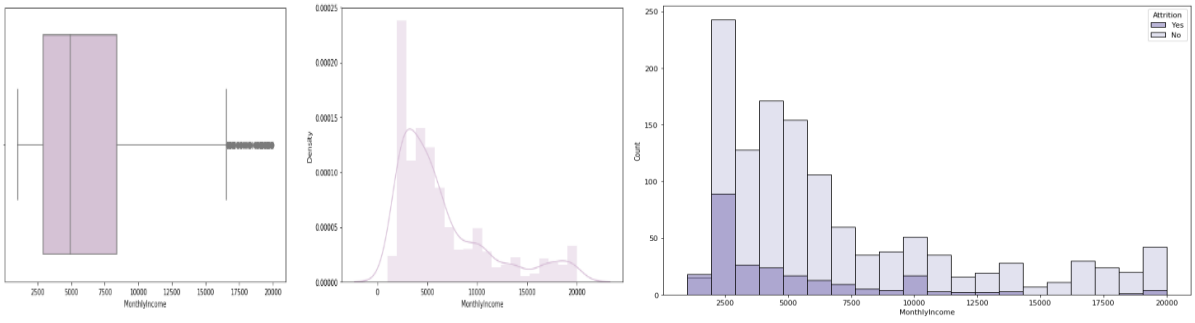
When it comes to the findings, the observation of satisfaction is simple because all three satisfaction variables indicate that low satisfaction is a factor in leaving an organization.

Summary

Findings from the investigation on work engagement. Before completely comprehending the influence of income on turnover, let's look at the link between pay and reward since pay and reward are usually the common thoughts of employees, which is worth studying. Companies should thus pay more attention to those workers who work harder but receive much less exchange. Such workers should receive more assistance, including training, job coaching, etc.; salary is frequently one of the rewards, even if they don't work as hard or know how to work properly.

2.4.7 Explore the relationship between variables such as income and input and employee turnover

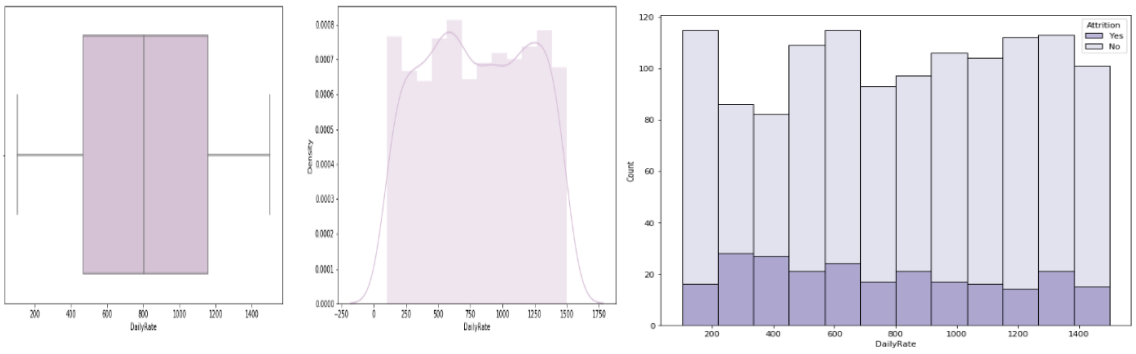
Monthly Income



Most employees make less than $10,000 per month and the median salary is $5,000. Through in-depth analysis and elaboration by coding, we observed there are 81 employees who make more than $17,500 monthly. The monthly salary for 200 employees is between $10,000 to $17,500. 1189 employees who make less than $10,000 monthly and 749 employees who

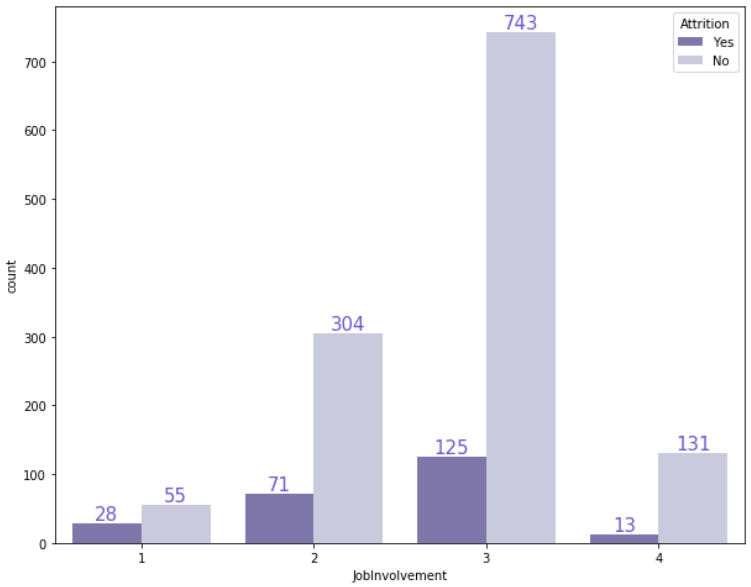
make less than $5,000 monthly. Lower-paid employees have a higher turnover rate than higher-paid employees, and the turnover rate for employees earning around $10,000 is also high.

Daily Rate



In terms of daily rate, most employees have 450 to less than 1200, with 800 being the median.

Job Involvement

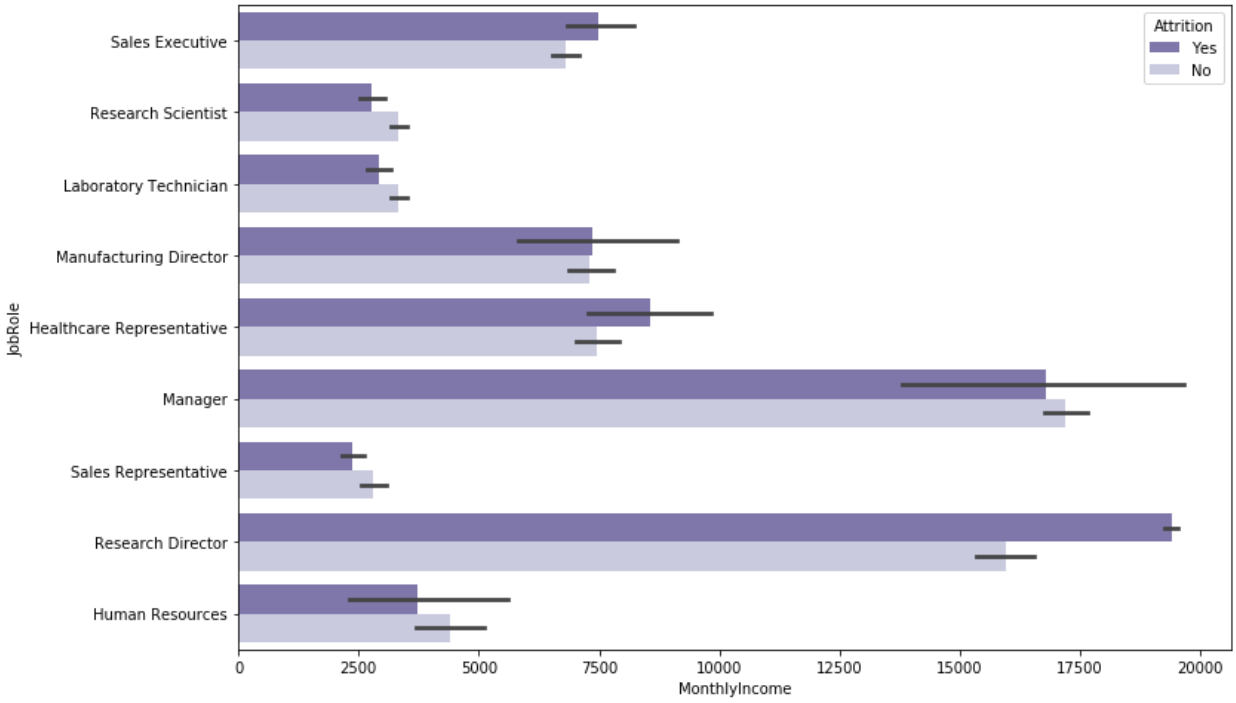


In terms of work involvement, the data visualization above shows that employees that are disengaged have a greater turnover rate. This is a very interesting conclusion for high or low income, which does not precisely show that low income is the cause of employee turnover, but here we can see that the wider the discrepancy between input and return, the more likely turnover that there will be.

Summary

The investigation of employee compensation revealed that low-income workers have a significant incidence of turnover. However, there is a modest peak in income at around $10,000, showing that employees at that level also have a high turnover rate. The company's top talent may have greater aspirations for them or various reasons to quit, making them the subject of attention, which is one explanation for this.

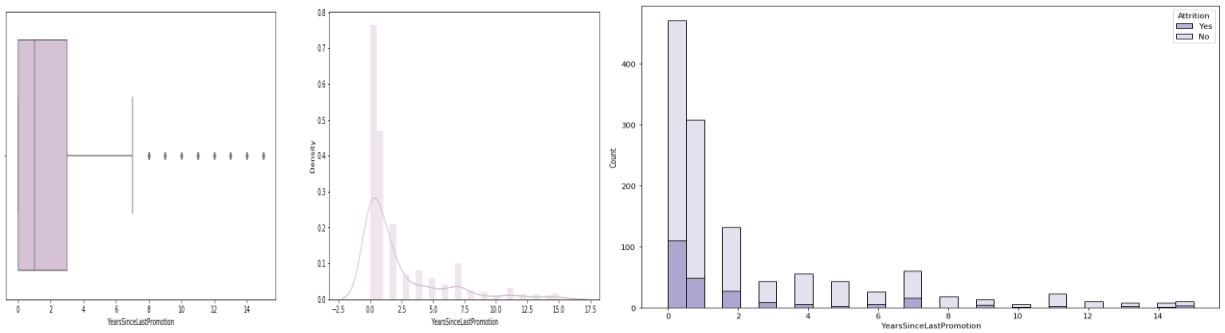
Does Job Role and Monthly Income affect attrition?



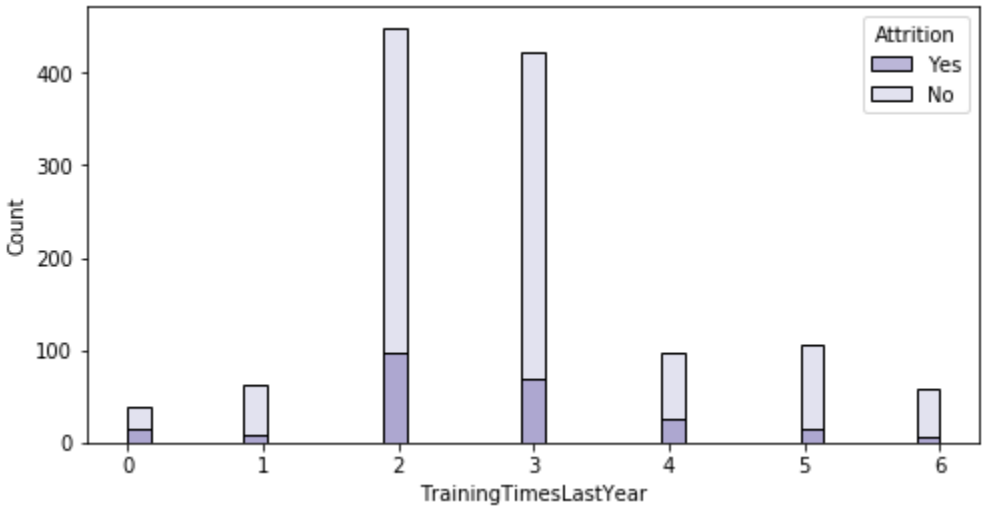
According to exploring the relationship between monthly income and job roles, we observed that the attrition flow of the Manager, Manufacturing Director, Sales executive, Laboratory Technician, Healthcare representative and Human resources are enormous. It is worth noting that the salaries of research scientists, laboratory technicians, and sales representatives are exceptionally low, which is the main reason in our consideration of turnover. In addition, Human resources department has the higher attrition when we look at the proportion during explosion on the Department variable. That might be a reason in our consideration of turnover.

2.4.8 Explore the relationship between variables such as employee promotions and employee turnover

Years Since Last Promotion



In this dataset, most employees have been promoted between 0-3 years since their last promotion, with a median of 1 year. Through in-depth analysis and coding elaboration, we found that 107 employees had been promoted more than 7 years since their last promotion. 425 employees had been promoted between 2-7 years since their last promotion.



Most employees have been promoted for less than 2 years, at 938. Even if the turnover rate for workers who didn't attend training the previous year is high, it isn't unusually high when compared to those who did.

3. Main Analysis

There are four models in the main analysis, namely Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine. The "Attrition" indicates whether the employee has left or not, if it is "1", the employee has left; if it is "0", the employee has not yet left. As "Attrition" our predictor variables.

In this machine learning research, the target variables will be divided into three groups in order to determine which group has the highest model performance and accuracy, as well as to determine which features and factors have the most impact on the prediction model.

Set 1. Use RFECV method and contains the ranked variables

Set 2. contains all variables except non-numerical and strongly correlated variables

Set 3. contains all variables and dummy all non-numerical variables

Summary Prediction Accuracy of All Three Test Sets

Test set 1	
Decision Tree	85.26%
Random Forest	87.53%
Logistic Regression	90.31%
SVM	84.35%

Test set 2	
Decision Tree	84.35%
Random Forest	87.30%
Logistic Regression	86.22%
SVM	84.35%

Test set 3	
Decision Tree	85.71%
Random Forest	86.85%
Logistic Regression	85.88%
SVM	84.35%

From the results, we can see that each test set has close accuracy for decision trees, random forests, and SVMs. The accuracy of decision tree is about 85%, random forest is 87% and SVM is 84%. However, our test set 1 has the highest accuracy in logistic regression with 90.31%. Therefore, this report will focus on reporting the highest model performance and accuracy for Test set1, and explain Random Forest, Decision Tree, and Logistic Regression models in detail.

3.1 Data Cleaning

Before machine learning, we examine the source dataset for NaN values and dimensionality reduction, and we discover that there are some meaningless variables, such as “Over18”, "EmployeeCount" ,"StandardHours,"and”EmployeeNumber” as each row variables contains information about one single employee who is adult, and each employee has the same

80 working hours and recorded with their unique employee ID number. Also, "Yes" and "No" are converted to “1” and “0”. This covers overtime and attrition, and as the original dataset contains only performance ratings of “3” and “4”, we will also convert to “0” and “1”.

3.2 RFECV

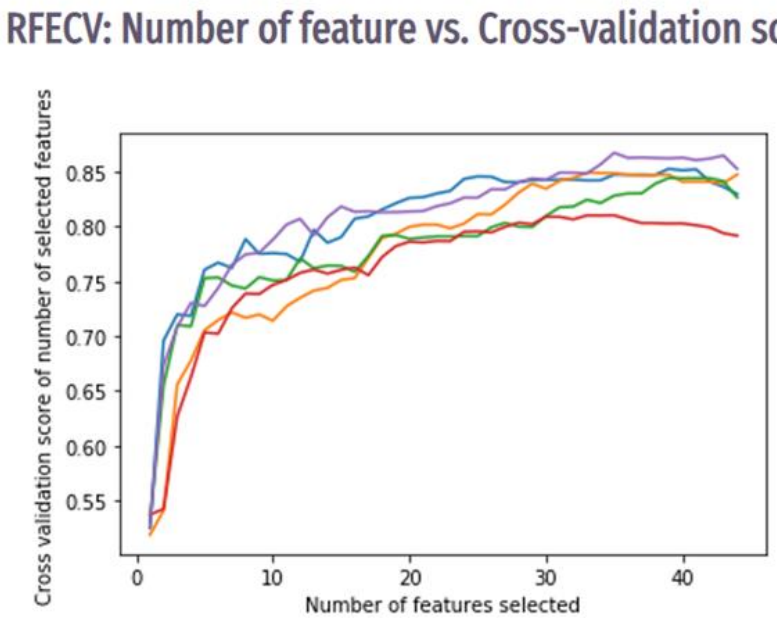
We use the RFECV approach which is known as recursive feature elimination cross validation. It is a feature selection method that optimizes a model by eliminating the weakest features until a certain number is attained. A limited number of features are deleted recursively by each loop, based on the coefficients of the model or the characteristic of feature significance.

We set the step=1 which means corresponds to the number of features to be removed at each iteration; cv = 5 as default 5-fold cross-validation; scoring = ‘roc_auc’ use sklearn scorer as input object; n_jobs =-1 using all processors (scikit-learn).

```
estimator_LR = LogisticRegression(C=2.1, penalty='l2',  solver='liblinear')
selector_LR = RFECV(estimator_LR, step=1, cv=5,scoring='roc_auc', n_jobs = -1)
selector_LR = selector_LR.fit(df, target)
print('Number of features :', selector_LR.n_features_)
print('Best features :', df.columns[selector_LR.support_])
```

```
Optimal number of features : 41
Best features : Index(['Age', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction',
       'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'NumCompaniesWorked',
       'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager', 'Gender_Male', 'JobRole_Human Resources',
       'JobRole_Laboratory Technician', 'JobRole_Manager',
       'JobRole_Manufacturing Director', 'JobRole_Research Director',
       'JobRole_Research Scientist', 'JobRole_Sales Executive',
       'JobRole_Sales Representative', 'BusinessTravel_Travel_Frequently',
       'BusinessTravel_Travel_Rarely', 'Department_Research & Development',
       'Department_Sales', 'EducationField_Life Sciences',
       'EducationField_Marketing', 'EducationField_Medical',
       'EducationField_Other', 'EducationField_Technical Degree',
       'MaritalStatus_Married', 'MaritalStatus_Single'],
      dtype='object')
```

From the figure we can see that the best score is obtained when there are more than 35 features, so we will keep 41 variables and use them as target variables



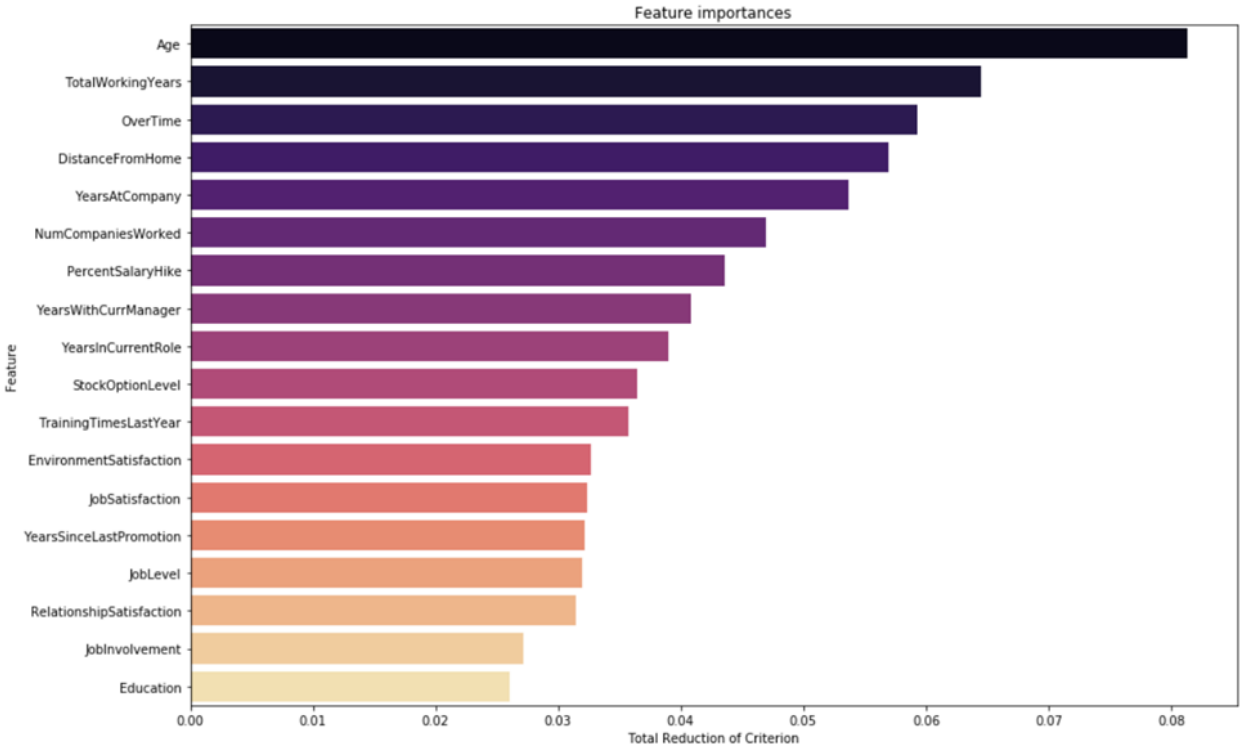
3.3 Random Forest

Using Random Forest we will get the ranking of the importance variables and will create random_forest of 10,100,250,500,750 and finally get the average value of 0.87

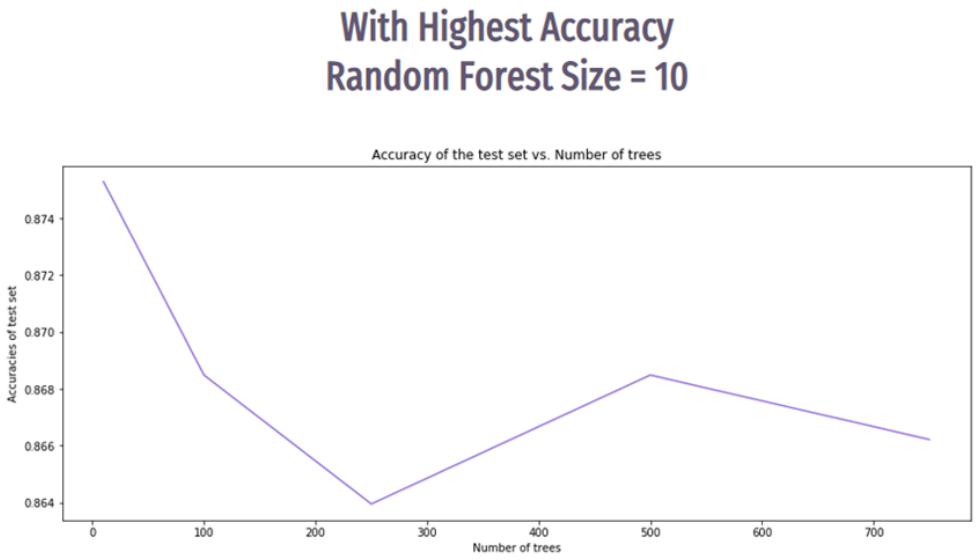
```
accuracies_rf4 = []
random_forest = [10,100,250,500,750]
for i in random_forest:
    model_rf = RandomForestClassifier(n_estimators=i, random_state=101)
    model_rf.fit(X_train, y_train)
    y_pred = model_rf.predict(X_test)
    confusion_matrix_results = confusion_matrix(y_test, y_pred)
    acc = display_metrics_acc('Random Forest', y_pred, y_test)
    accuracies_rf4.append(acc)

accuracies_rf4
```

```
[0.8752834467120182,
 0.8684807256235828,
 0.8639455782312925,
 0.8684807256235828,
 0.8662131519274376]
```



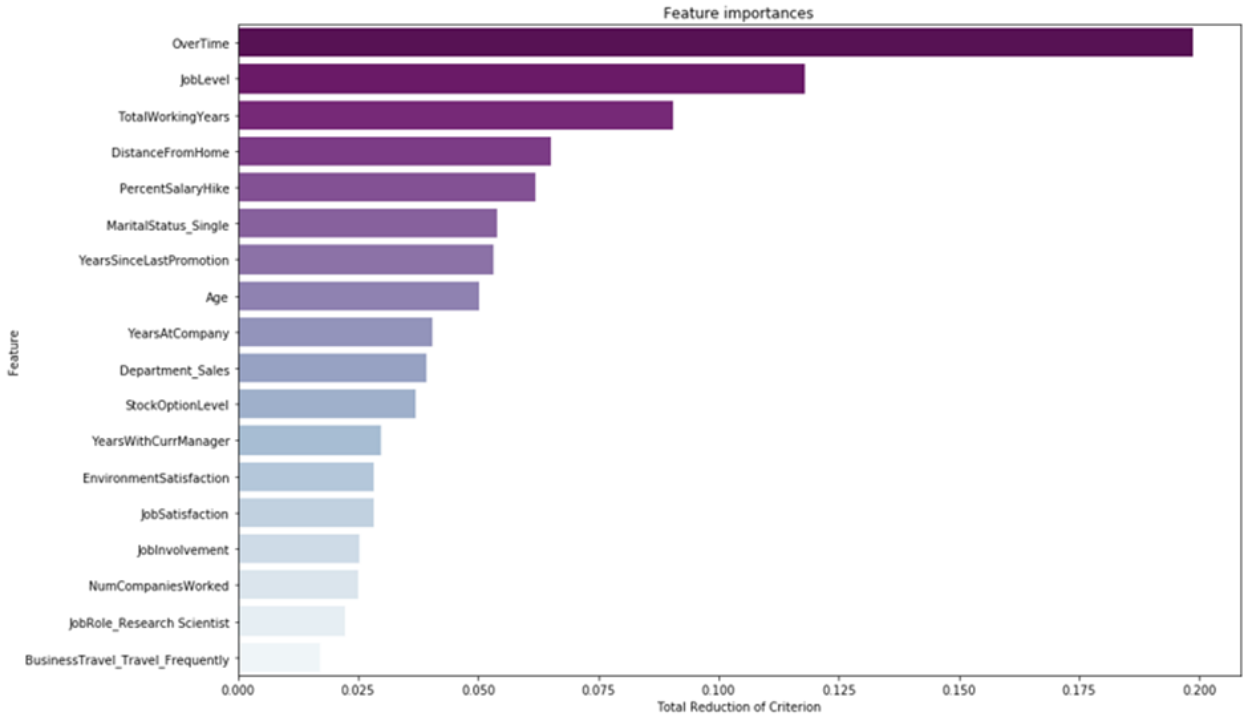
Using Random Forest model we get Top 5 Variable Importance: Age, Total Working Years, Over Time, Distance From Home, Year At Company. We can get that age is also an important factor and we can combine it with EDA tables to find why 30-year-old employees have a higher turnover rate. We also used the for loop to check which random forest size has highest accuracy score which is 10



3.4 Decision Tree

From Decision Tree feature ranking we can get the top 5 variables or the most important factors that affect attrition. They are Over Time, Job Level, Total Working Years, Distance from Home, Percent Salary Hike. We can see that overtime makes employees more

stressed, which leads to a greater tendency to quit the job.



We also used the for loop max_depth size with 5,6,7,8,9 to check which one has highest accuracy score which is 5.0. And average with 0.83 accuracy.

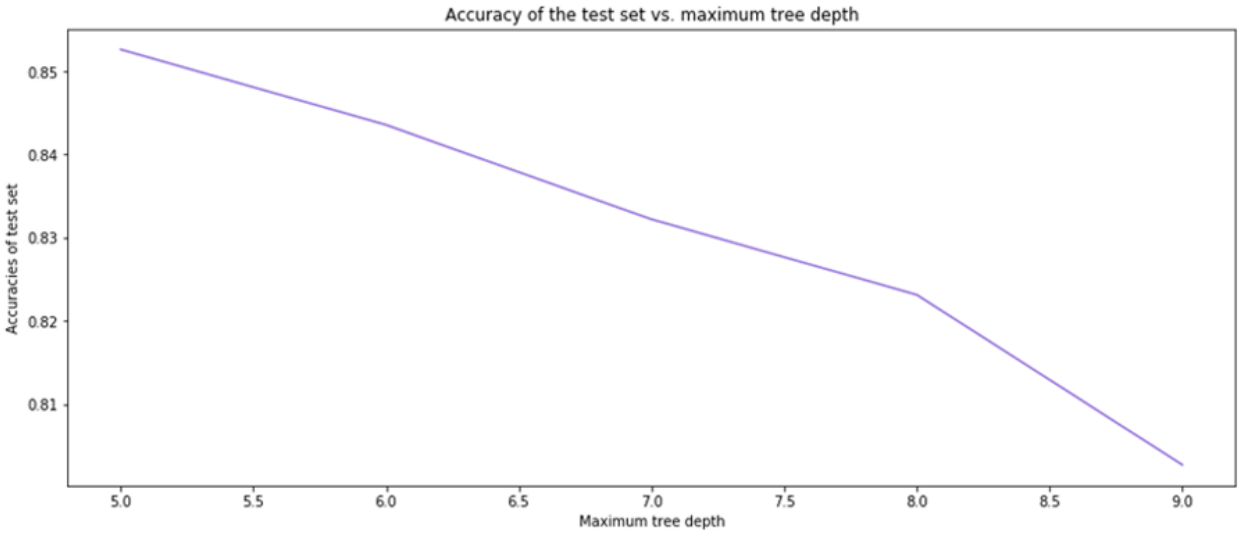
```
accuracies_dt4 = []
max_depth = [5,6,7,8,9]
for i in max_depth:
    model_dt = DecisionTreeClassifier(criterion='entropy', random_state=42, max_depth=i)
    model_dt.fit(X_train, y_train)
    y_pred = model_dt.predict(X_test)
    acc = display_metrics_acc('Decision Tree',y_pred,y_test)

    accuracies_dt4.append(acc)

accuracies_dt4
```

```
[0.8526077097505669,
 0.8435374149659864,
 0.8321995464852607,
 0.8231292517006803,
 0.8027210884353742]
```

With Highest Accuracy
Tree depth = 5.0



3.5 Logistic Regression

we can use sklearn and statemodel with summary which includes coefficient and p value and use p value to find if the predictor variables have a statistically significant relationship with the target variable in the model.

	coef	std err	z	P>|z|	[0.025	0.975]
Age	-0.0313	0.014	-2.319	0.020	-0.058	-0.005
DistanceFromHome	0.0458	0.011	4.271	0.000	0.025	0.067
Education	0.0034	0.087	0.039	0.969	-0.168	0.175
EnvironmentSatisfaction	-0.4334	0.083	-5.242	0.000	-0.595	-0.271
JobInvolvement	-0.5314	0.122	-4.348	0.000	-0.771	-0.292
JobLevel	-0.0543	0.231	-0.235	0.814	-0.507	0.398
JobSatisfaction	-0.4186	0.081	-5.156	0.000	-0.578	-0.259
NumCompaniesWorked	0.1937	0.039	5.013	0.000	0.118	0.269
OverTime	1.9705	0.193	10.214	0.000	1.592	2.349
PercentSalaryHike	-0.0218	0.039	-0.557	0.577	-0.098	0.055
PerformanceRating	0.1041	0.396	0.263	0.792	-0.671	0.880
RelationshipSatisfaction	-0.2571	0.082	-3.121	0.002	-0.419	-0.096
StandardHours	-0.2510	1063.884	-0.000	1.000	-2085.425	2084.923
StockOptionLevel	-0.2088	0.157	-1.332	0.183	-0.516	0.098
TotalWorkingYears	-0.0610	0.029	-2.087	0.037	-0.118	-0.004
TrainingTimesLastYear	-0.1920	0.073	-2.630	0.009	-0.335	-0.049
WorkLifeBalance	-0.3632	0.123	-2.943	0.003	-0.605	-0.121
YearsAtCompany	0.0945	0.039	2.429	0.015	0.018	0.171
YearsInCurrentRole	-0.1516	0.045	-3.356	0.001	-0.240	-0.063
YearsSinceLastPromotion	0.1782	0.042	4.244	0.000	0.096	0.261
YearsWithCurrManager	-0.1350	0.047	-2.878	0.004	-0.227	-0.043
Gender_Male	0.3884	0.184	2.112	0.035	0.028	0.749

As result, any variable has p value that less than 0.05 can consider as statistically significant relationship. They are age, relationship satisfaction, total working years, training time last year, work life balance, years at company, years in current role, years with current manager. For example, p value of Years at company is 0.015 which is less than 0.05 so it has a statistically significant relationship with the response variable in the model. On the contrary, Education is 0.969 and greater than 0.05 which does not have strong relationship.

Moreover, if employee's distance from home is 1 unit more, his/ her chance of leaving the company will increase by 0.0458 units. And if over time is 1 unit more, his/her chance of

leaving the company will increased by 1.9705 which is the highest value. In terms of output interpretability, logistic regression provides a single value for each predictor variable, as well as a confidence interval. Random Forest and Decision trees give clear classification rules and feature ranking that helps audience to indicate the main factors.

In conclusion, Logistic Regression is the most used technique for addressing issues on an industrial scale. In terms of output interpretation, you will get the value or interval of the association between each predictor variable and the target variable, which you may then utilize as necessary. The benefit is that it is simple and straightforward to work with, as well as very efficient in terms of processing and memory needs. While Decision Tree and Random Forest will provide a clear ranking of the features and help the audience understand the most important factors in terms of conclusions, the model will become complex if there are more trees, and less accurate and inefficient models will be produced before cross-validation is implemented.

4. Findings & Recommendation

4.1 Organization Based

After determining the causes of employee departure, we may assist the organization in making decisions based on the model and its outcomes.

1) According to research, increasing pay has a major influence on how much workers value their percentage compensation rises and if they opt to remain with the company.
2) Regularly give workers with promotion opportunities so they feel appreciated, particularly loyal employees who have been with the firm for an extended period of

time, to encourage and foster their entrepreneurship. Regularly engaging with senior employees to channel their promotion-related discontent in a timely way. Internal promotions should be timely and appropriate. To prevent senior personnel from leaving the organization, especially at the management level, higher-level leaders must communicate effectively in advance of an organization reorganization.

3) The majority of workers reside more than 15 miles away from the office; thus, HR must inform candidates during the interview process. If the distance between the house and the job is great, workers will spend more time traveling; candidates may consider renting an apartment near to the workplace. Human Resources must evaluate the risk associated with picking the individual if they answer that they can adapt despite the great distance. Because if an employee commutes for more than two hours per day for more than a week or a month, this is acceptable in the short term. However, if this situation persists for an extended period of time, the employee will be more likely to depart due to other considerations.
4) Allowing workers to work excessive overtime produces an increase in weariness and has negative effects on job satisfaction and turnover.

4.2 Data Based

In terms of future employee data collection, firms may give the following factors to increase the accuracy of prediction models and to provide the company with more thorough employee profiles.

1) The source data does not indicate whether a particular employee quit or was terminated.
2) The source data does not include the date of starting and leaving the company; according to the research, turnover rate refers to the percentage of employees who have left the company within a given time period; and it is impossible to obtain more accurate data due to the lack of information on the data.
3) The previous data of workers who left the organization is a tiny proportion; if we can get more information about employees who leave, our projection will be more accurate.

References

Sklearn.feature_selection.RFECV. scikit. (n.d.). Retrieved November 21, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html