Project 1: HR Analytics and Prediction of Employee Attrition

Kesav Adithya Venkidusamy

Bellevue University - Master of Science in Data Science

DSC680-T301 Applied Data Science (2231-1)

Professor Catherine Williams

09/25/2022

Table of Contents

Business Problem

Employee retention strategies are integral to the success and well-being of a company. Attrition is a problem that impacts all businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff. As such, there is great business interest in understanding the drivers of, and minimizing staff attrition. There are often many reasons why employees leave an organization, and in this case study, I will explore some of the key drivers of employee attrition.

Background/History

Employee attrition measures how many workers have left an organization and is a common metric companies use to assess their performance. Some of the common reasons for employee attrition are as follows.

1. Poor job satisfaction and pay
2. Not enough career opportunity
3. Poor workplace culture
4. Lack of employee motivation
5. Poor work-life balance
6. Not fitting in and feeling sense of belonging

While turnover rates vary from industry to industry, the Bureau of Labor Statistics reported that among voluntary separations the overall turnover rate was 32.7% in 2021, and even more than this in 2022. So, predictive attrition model helps in not only taking preventive measures but also into making better hiring decisions. Minimizing attrition can ensure associates stay longer, enabling them to continue benefiting the organization operations.

Exploratory Data Analysis

The dataset is extracted from the following Kaggle website. This is a fictional dataset created by IBM data scientists. The dataset contains approximately 1500 entries.

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

This dataset presents an employee survey from IBM, indicating if there is attrition or not. Using this dataset, I will uncover the factors that lead to employee attrition and explore some of factors contribute to the attritions.

Characteristics

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer
Associated Tasks	Classification
Number of Instances	1470
Number of Attributes	35
Missing Values	No
Area	Social

Data Dictionary

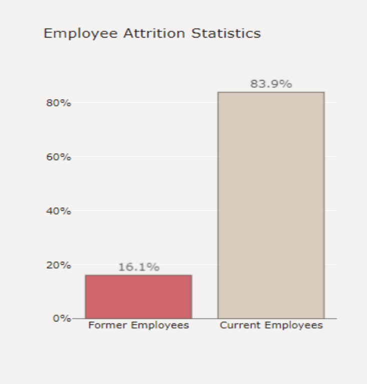
Feature Name	**Feature Description**	**Feature Type**
Age	Age of the person	Continuous
Attrition	Person has left the company or not	Target
BusinessTravel	How frequently the person travels	Discrete
DailyRate	Daily Rate for the employee	Continuous
Department	Department of the person	Discrete
DistanceFromHome	Distance of the company from home	Continuous
Education	Education of the person 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'	Discrete
EducationField	Education field of the person	Discrete
EmployeeCount	Count of Employee	Discrete
EmployeeNumber	Employee id of the person	Continuous
EnvironmentSatisfaction	Environment Satisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Discrete
Gender	Gender	Discrete
HourlyRate	Hourly rate for the employee	Continuous
JobInvolvement	Involvement in job 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Discrete

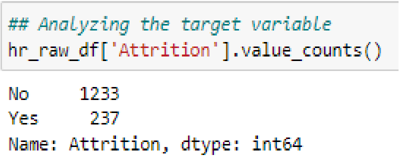
JobLevel	Job Level	Discrete
JobRole	Job Role	Discrete
JobSatisfaction	Job Satisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'	Discrete
MaritalStatus	Marital Status of Employee	Discrete
MonthlyIncome	Monthly Income of the person	Continuous
MonthlyRate	Monthly Rate	Continuous
NumCompaniesWorked	Number of Companies worked	Discrete
Over18	Over 18 years	Discrete
OverTime	Worked over time	Discrete
PercentSalaryHike	Percentage of Salary Hike	Continuous
PerformanceRating	Performance Rating 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'	Discrete
RelationshipSatisfaction	Relationship Satisfaction for the employee 1 'Low' 2 'Medium' 3 'High' 4 'Very High	Discrete
StandardHours	Standard work hours	Discrete
StockOptionLevel	Stock Option Level given to the employee	Discrete
TotalWorkingYears	Total Number of years worked	Continuous
TrainingTimesLastYear	Training times attended during last year	Continuous
WorkLifeBalance	Work Life Balance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'	Discrete
YearsAtCompany	Number of years with current company	Continuous
YearsInCurrentRole	Number of years in the current role	Continuous
YearsSinceLastPromotion	Number of years since last promotion	Continuous
YearsWithCurrManager	Number of years with current manager	Continuous

Data Preparation

The problem statement of this project is to identify the dataset feature(s) which are mostly related to or affecting the employee attrition rate. The data set contains approximately 1500 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs a random allocation of probability of attrition. The dataset consists of 34 features of which 26 are numerical and rest all are categorical with “attrition” being the target.

The target variable “attrition” contains 2 values “Yes” and “No” which would be subsequently converted to 1 and 0 respectively. The percentage of these values are 16.1% and 83.9% respectively, suggesting the employees who left the organization (Attrition = “Yes”) are significantly less compared to those who are continuing in the organization (Attrition = “No”), making our data set kind of imbalanced considering the target variable. The details are shown in figure 2.





```
## Analyzing the target variable
hr_raw_df['Attrition'].value_counts()

No     1233
Yes     237
Name: Attrition, dtype: int64
```

Figure 1. Target Variable “Attrition” Analysis

As part of EDA process, null check has been performed initially on all the variable fields and found no null value being present in any of the feature. Then, duplicate check has also been performed and found no duplicate in the dataset. Finally, few features as mentioned below are removed from the dataset as they do not add any value-add to the target variable “attrition”.

1. EmployeeCount - constant value "1"
2. StandardHours - constant value "80"
3. Over18 - constant value "True"
4. EmployeeNumber - Employee number is key column which can also be removed
5. StockOptionLevel - I believe this is stock options given to the employees having values between 1 to 3;

Data Visualization

As mentioned before, the dataset contains 26 numerical features and 8 categorical features, and ‘attrition feature being the target variable. After removing the unwanted features, below is the final list for numerical and categorical variables.

Histogram

Histogram is used to identify the distribution of numerical features present in the dataset. Among various numerical features, only 12 significant features are considered for histogram as shown in figure 3.

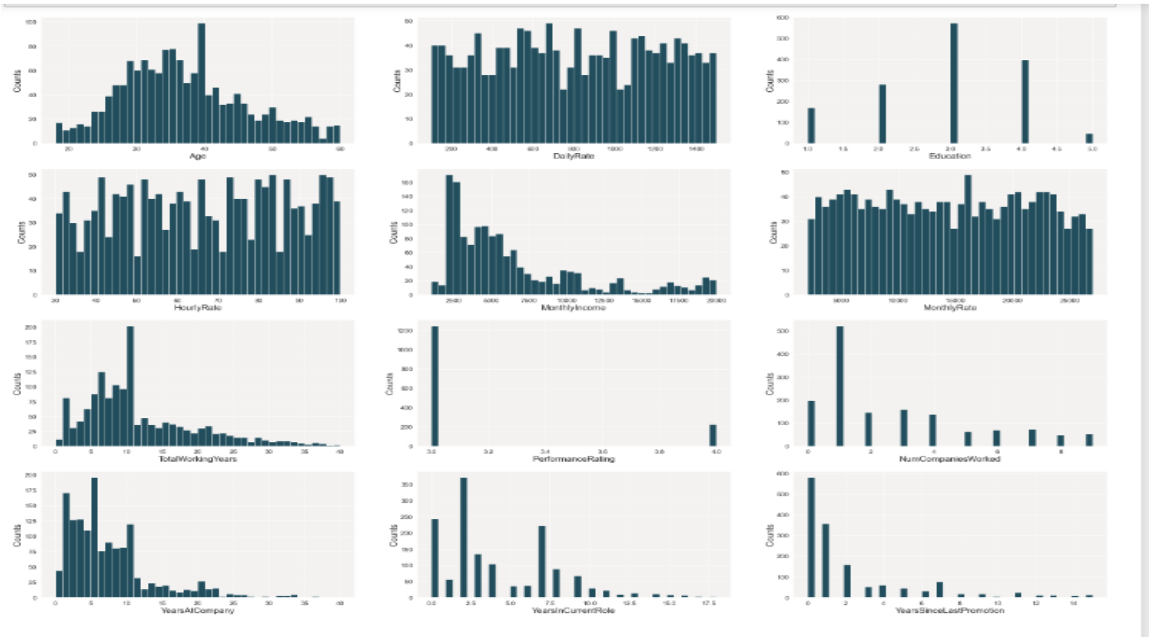
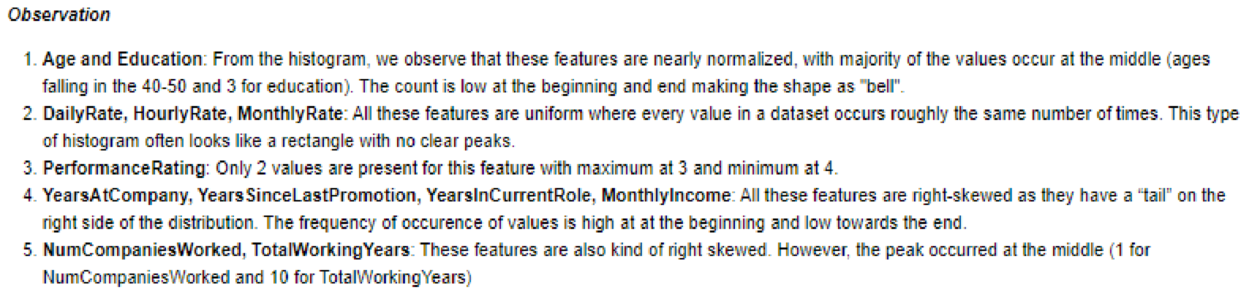


Figure 2 Histogram for numerical variables



Observation

1. **Age and Education**: From the histogram, we observe that these features are nearly normalized, with majority of the values occur at the middle (ages falling in the 40-50 and 3 for education). The count is low at the beginning and end making the shape as "bell".
2. **DailyRate, HourlyRate, MonthlyRate**: All these features are uniform where every value in a dataset occurs roughly the same number of times. This type of histogram often looks like a rectangle with no clear peaks.
3. **PerformanceRating**: Only 2 values are present for this feature with maximum at 3 and minimum at 4.
4. **YearsAtCompany, YearsSinceLastPromotion, YearsInCurrentRole, MonthlyIncome**: All these features are right-skewed as they have a "tail" on the right side of the distribution. The frequency of occurence of values is high at at the beginning and low towards the end.
5. **NumCompaniesWorked, TotalWorkingYears**: These features are also kind of right skewed. However, the peak occurred at the middle (1 for NumCompaniesWorked and 10 for TotalWorkingYears)

Bar Graph

Bar graph has been plotted for all categorical features to understand the distribution of data among unique values. The details are shown in figure 4.

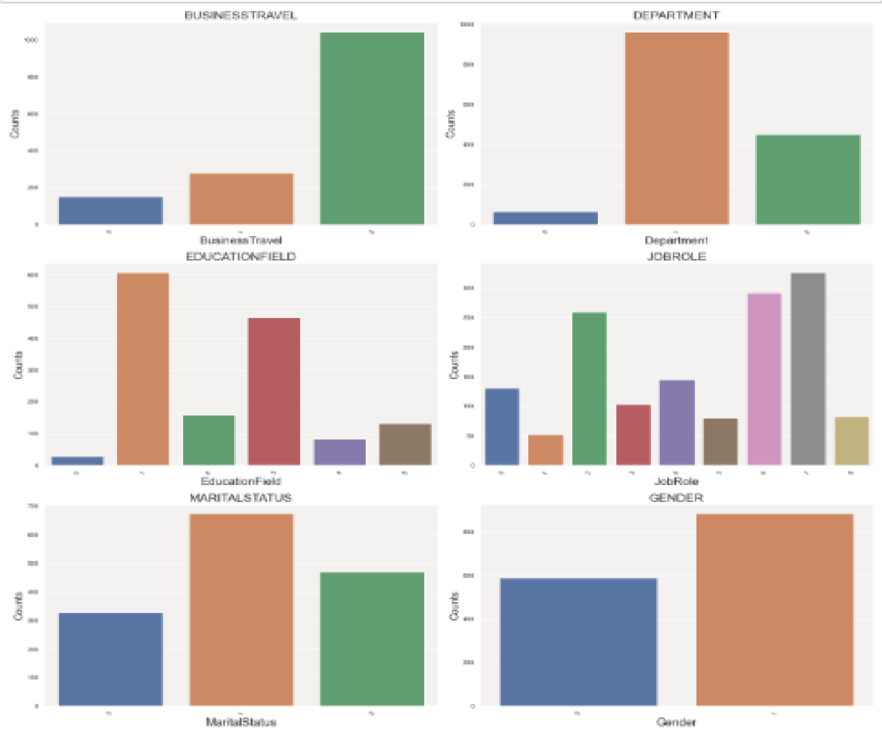


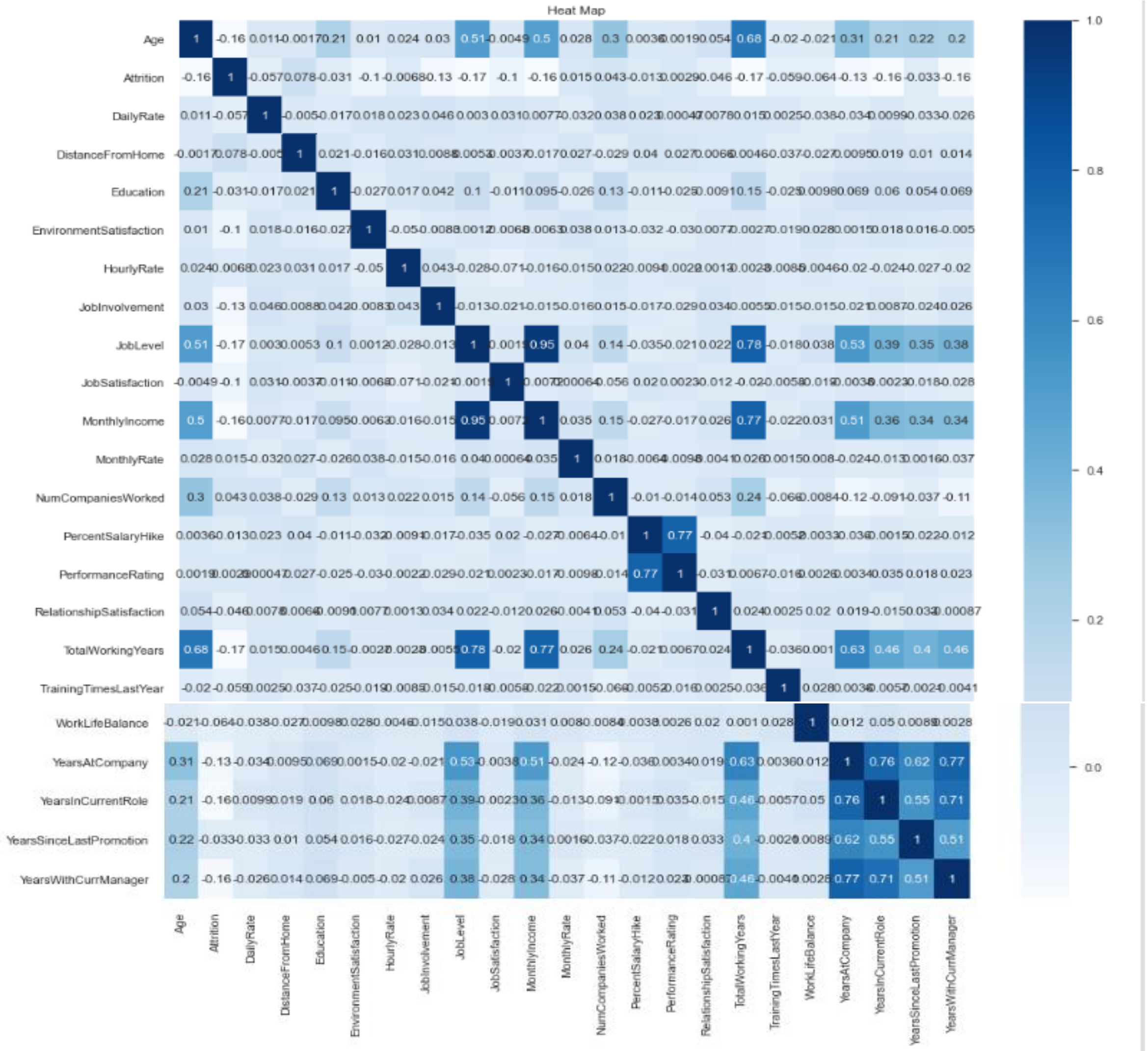
Figure 3 Bar Graph for Categorical Variables

Observation

1. BusinessTravel - There are 3 unique values present for this feature with more number of records (> 1000) for "Travel_Rarely" and slightly greater than 200 for Travel_Frequently and slightly less than 200 for Non-Travel.
2. Department - This feature has 3 unique values with more number of records for R&D Department and least number of records for HR department
3. Education - This feature has 6 different values with more people falling under Life Science and Medical and less number of people HR and Other
4. JobRole - There are many values present for this feature with more people falling under SalesExecutive, Research Scientist, Laboratory Technician and less people under the job roles sales representative, research director and HR.
5. MaritalStatus - Marital status field has 3 distinct values with most of the people falling under marries and less number of people under Divorced.
6. Gender - Final categorical variables is Gender with 2 distinct values Males and Female with more number of records for Males compared to Female.

Heat Map

Heat map has been created to understand correlation between various features present in the dataset. The details are shown in the figure 5.



Observation

Confirming our findings in the scatterplot above, MonthlyIncome has a strong positive correlation to TotalWorkingYears of 0.77. Additionally, YearsAtCompany has a strong positive association with YearsWithCurrManager (correlation = 0.77), as well as with YearsInCurrentRole (correlation = 0.76).

Monthly income is also having strong correlation with JobLevel of 0.95 which makes complete sense as Monthly income increase for increase in Job Level.

Age is having postive correlation of 0.68 with TotalWorkingYears which also make complete sense. For increase in age, TotalWorkingYears will also increase.

The vriables YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager are all postively correlated with each other.

Performance Rating variable is positively correlated (~0.77) with Percent Salary Hike. For increase in Performance rating, the in Percenta Salary hike will also increase

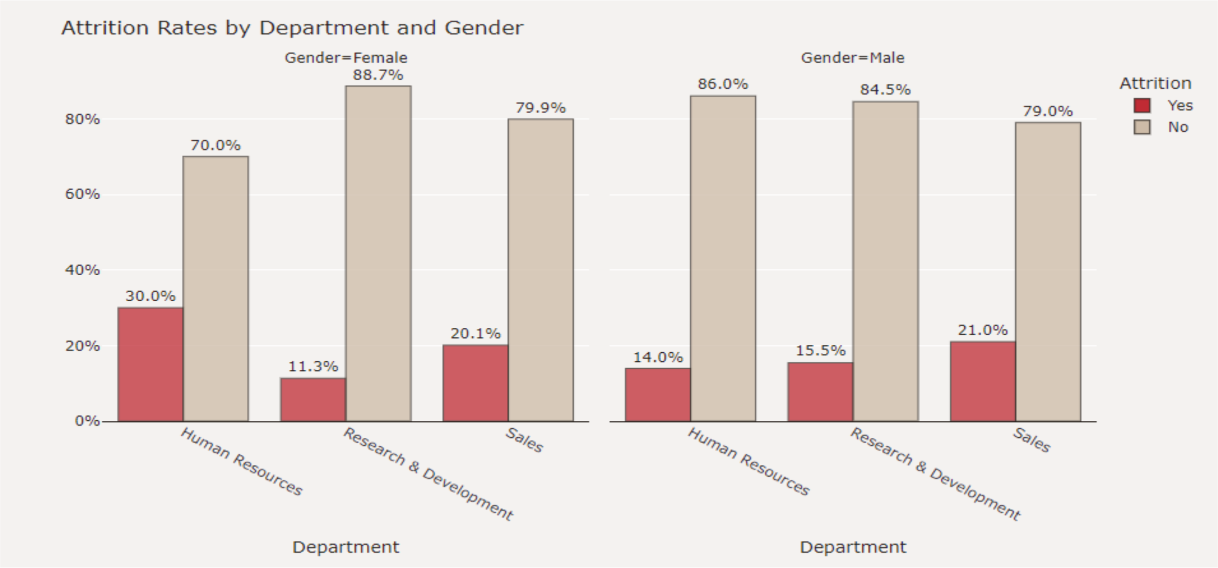
Attrition variable is having positive correlation with DistanceFromHome, NumCompaniesWorked, JobSatisfaction, JobLevel and HourlyRate

Figure 4 Heat Map Correlations and observation

Target Variable Analysis

Initially, grouped bar chart has been created to compare attrition rate “Yes” (represented as 1) and attrition rate “No” (represented as 0) for the categorical features. The details are shown in figure 6. Some of the observations found during grouped bar analysis are as follows.

- Women in Human Resources experienced the highest amount of turnover, with nearly 1 out of every 3 women in HR leaving the company. Sales department comes 2nd with turnover rate of 21% and Research & Development comes third with a rate of 11.3%.
- For men, the highest turnover occurred in the Sales department with nearly 21%. The remaining departments Research and Development and HR are more or less experiencing similar rate of attrition (~15%).
- Among women with the highest rated work life balance, 1 out of 4 left the company, the highest proportion among the ratings for women. For men, the highest proportion occurred in those with the lowest work life balance.
- Across each department, the average salary for women is higher than average salary of men
- In comparison to current employees, former employees had lower median salaries across all three departments. In Human Resources and Research and Development departments, women tend to have higher median salaries than men.



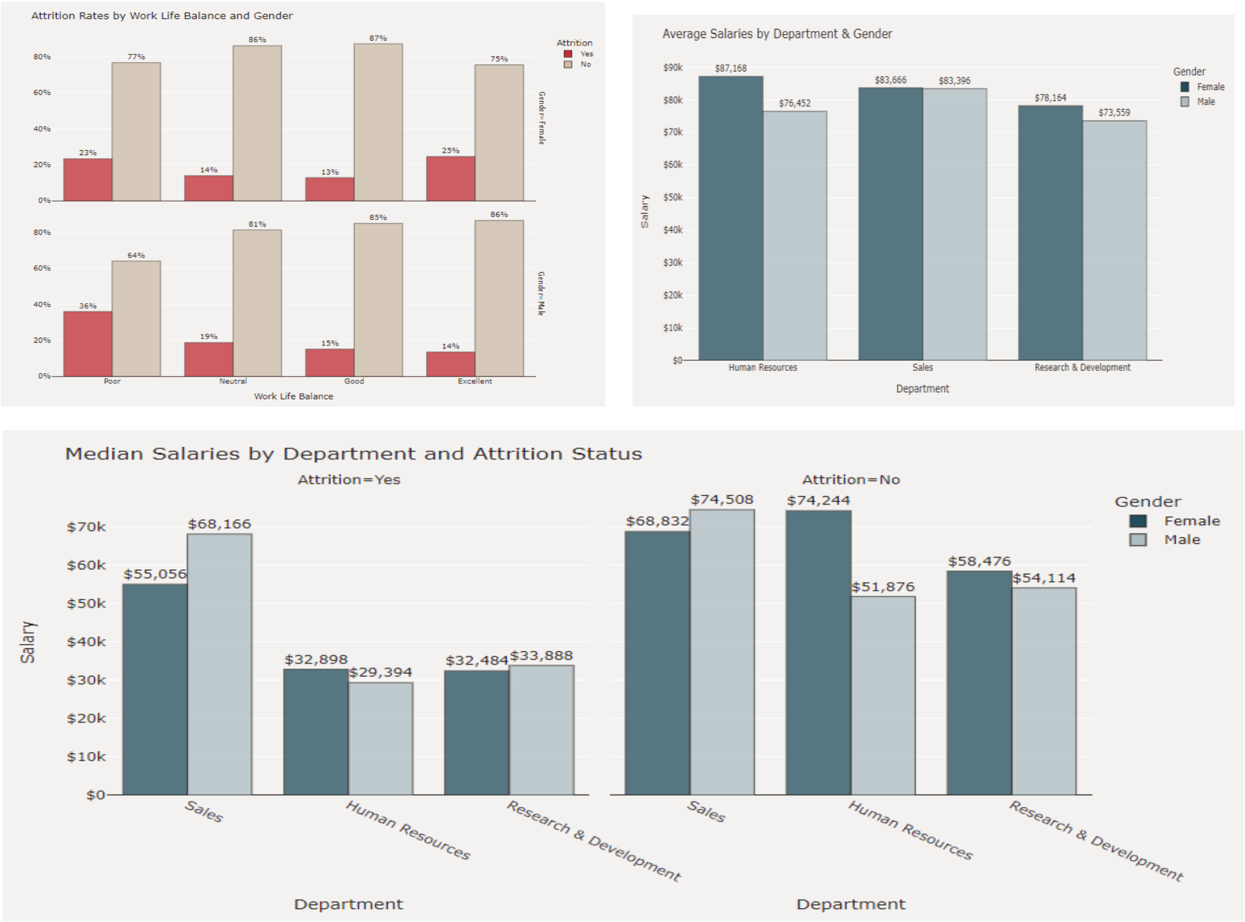


Figure 5 Grouped Bar Charts

Box Plot

A box plot is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum. Box charts have been plotted to show monthly income by work life balance for Attrition and monthly income by job role for attrition. The details are shown in the figure 7. Following are the observations out of box plot.

- The Attrition rate is high among the people who have "Good" work life balance earning median monthly income of 3202 dollars. However, the attrition rate is less among the people who have "Excellent" work life balance earning median monthly income of 2785 dollars; The people with "Fair" work life balance comes next.
- To a surprise, the attrition rate for the people who have "Poor" work life balance is less. We will analyze more for the reason.
- More number of people who have "Excellent" work life balance are continuing in the company which is as expected (Attrition Rate = No);

- The attrition rate is high among "Managers", "Manufacturing Director" and "Sales Executives" whereas it is low for "Research Director" and "Sales Representative"; So, "Research Director" and "Sales Representative" are not willing to quit the company often whereas "Managers", "Manufacturing Director" and "Sales Executives" often change their companies.

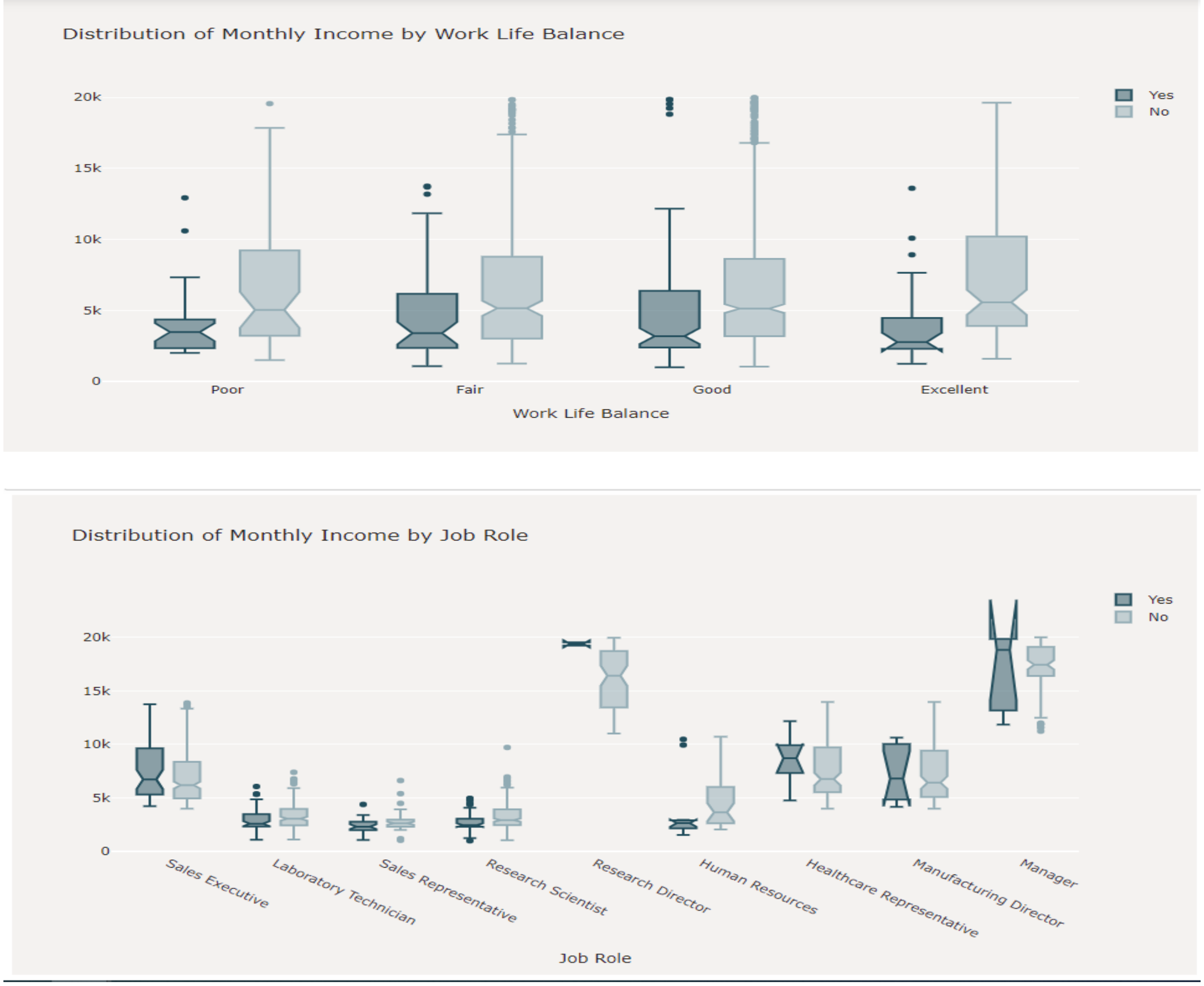


Figure 6 Box Plots

Scatter Plot

A scatterplot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. In the scatter plot, I analyzed the monthly income with total number of years worked and job level as shown in figure 7. I observed that monthly income is positively correlated with total number of years worked and there is strong association between an employee's earnings and their job level.

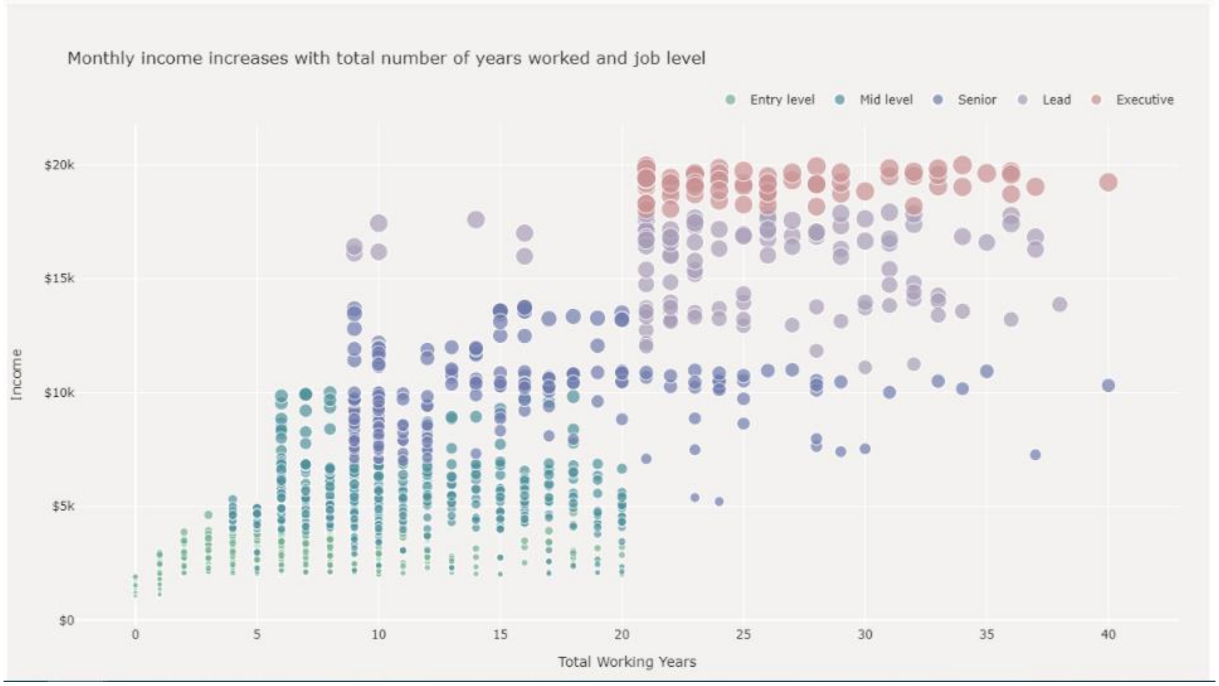


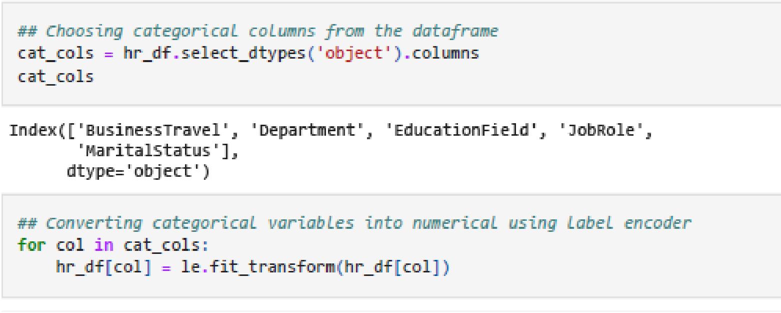
Figure 7 Scatter Plot for Monthly income increase with total number of years and job level

Feature Engineering

The following techniques have been applied on the dataset applying modeling techniques.

1. Label Encoder
2. SMOTE

Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering. Label encoder is applied on all the categorical values to convert into integer as shown in the figure 8.



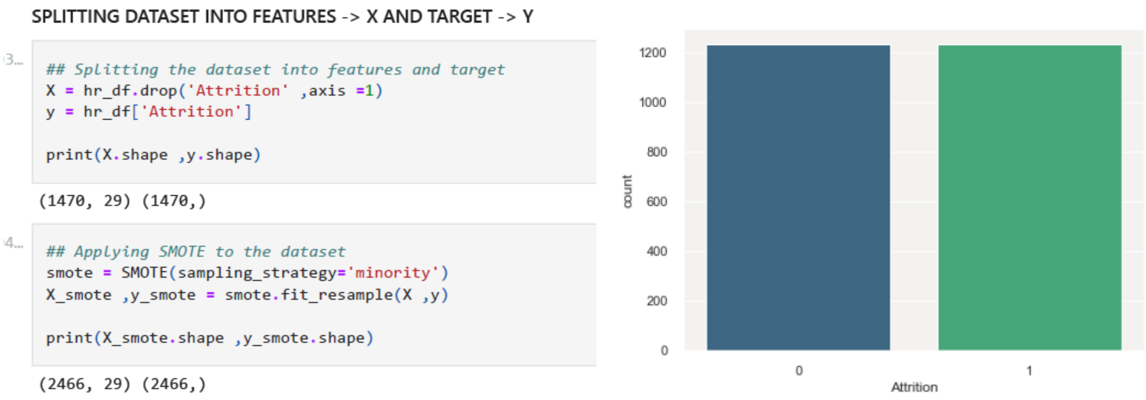
```
## Choosing categorical columns from the dataframe
cat_cols = hr_df.select_dtypes('object').columns
cat_cols
```

```
Index(['BusinessTravel', 'Department', 'EducationField', 'JobRole',
       'MaritalStatus'],
      dtype='object')
```

```
## Converting categorical variables into numerical using label encoder
for col in cat_cols:
    hr_df[col] = le.fit_transform(hr_df[col])
```

Figure 8 Label Encoder to convert categorical variables into numerical type

Since the values present in the target variable are extremely unbalanced, “SMOTE” method is leveraged to balance the data. SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. The details are shown in figure 9.



SPLITTING DATASET INTO FEATURES -> X AND TARGET -> Y

```
## Splitting the dataset into features and target
X = hr_df.drop('Attrition' ,axis =1)
y = hr_df['Attrition']

print(X.shape ,y.shape)
```

```
(1470, 29) (1470,)
```

```
## Applying SMOTE to the dataset
smote = SMOTE(sampling_strategy='minority')
X_smote ,y_smote = smote.fit_resample(X ,y)

print(X_smote.shape ,y_smote.shape)
```

```
(2466, 29) (2466,)
```

Figure 9 SMOTE method on the target variable “Attrition” to balance the data

Methods

The following modeling techniques are used to determine which modeling technique works best on this dataset and the features that are mostly related or correlated to our target variable “Attrition”.

1. Logistic Regression
2. Decision Tree
3. Random Forest

The accuracy and score for all the three methods have been calculated and compared to find out the best modelling technique for this dataset. I have also applied the standard scalar method/technique to the train and test data and applied the same model choices. Those results helped in deciding best predictive model.

Finally, Pearson correlation matrix, chi-squared, random forest classifier methods were used to determine the top 5 features which are mostly correlated with the target variable “Attrition”.

Analysis

Modeling Analysis

The scores that I received for each model are represented in the table below.

Model	Accuracy	F1 Score (Attrition = “Yes”)	F1 Score (Attrition = “No”)	AUC Score
Logistic Regression	65.27%	0.65	0.66	0.71
Decision Tree	82.16%	0.83	0.82	0.83
Random Forest	91.08%	0.91	0.91	0.97
Logistic Regression with Standard Scalar	81.62%	0.82	0.82	
Decision Tree with Standard Scalar	82.16%	0.83	0.82	
Random Forest with Standard Scalar	90.81%	0.91	0.91	
Logistic Regression – Target variable with 5 best features using X2	65.54%	0.66	0.65	0.70
Decision Tree – Target variable with 5 best features using X2	72.83%	0.74	0.71	0.73
Random Forest – Target variable with 5 best features using X2	80.81%	0.81	0.79	0.88

Accuracy: Accuracy represents the number of correctly classified data instance over the total number of data instances.

F1 Score: F1-Score is a metric which considers both precision and recall.

- **Precision:** Positive predictive value
- **Recall:** true positive rate

AUC Score: What area under the ROC curve describes good discrimination? We will use the following rule of thumb

- 0.5: This suggests no discrimination, so we might as well flip coin
- 0.5-0.7: We consider this as poor discrimination, not much better than a coin toss
- 0.7-0.8: Acceptable discrimination
- 0.8-0.9: Excellent discrimination
- >0.9: Outstanding discrimination

Among all 3 models and multiple iterations, we noticed the AUC score is high for the Random Forest which is 0.97 when we run on the dataset after removing unwanted features and without Standard Scalar. However, the score got reduced to 0.87 when we run on the dataset with only top 5 features. Applying standardization method (StandardScalar) on the dataset didn't improve the score much for decision tree and random forest classifier. Figure 10 represents a "Scores Plot", that shows the accuracy results of the different scenarios compiled in our analysis.

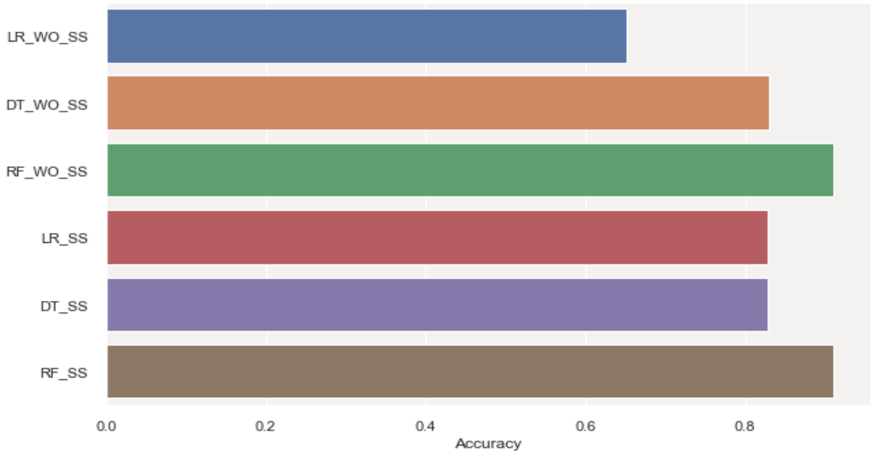


Figure 10: Score comparison across models

Further analysis has been done by creating confusion matrix for each model as depicted in figure 11. Logistic regression has high false negative value of 125 and low true positive value of 246 compared to decision tree classifier whose false negative and true positive values are 57 and 314 respectively. Random forest classifier shows further improvement where false negative is only 38 and true positive value is 333.

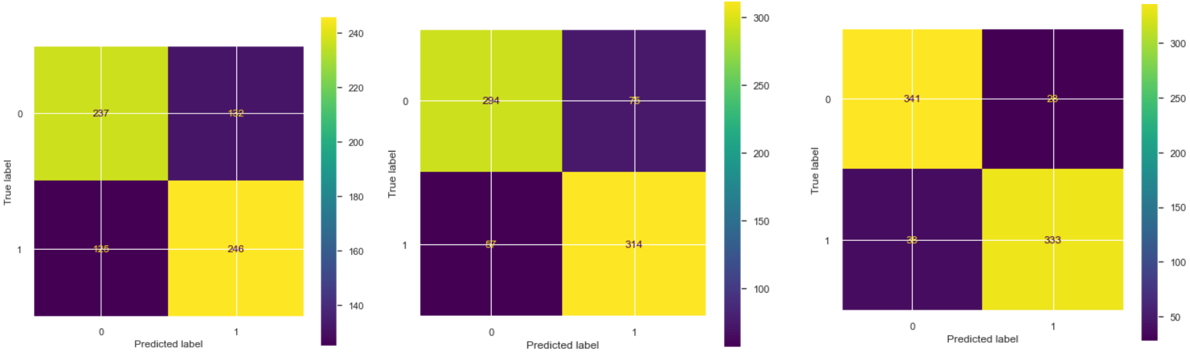


Figure 11: Confusion Matrix Plot for Logistic Regression, Decision Tree and Random Forest

Feature Analysis

Using various methods, I tried to find the best features from the dataset and following are the best features in the dataset which shows higher impact to the target variable "Attrition" compared to other features present in the dataset.

Test	Features
Pearson's correlation matrix - Feature correlation to target variable "income"	<u>**Positive Correlation**</u>**:** • OverTime • MaritalStatus • DistanceFromHome <u>**Negative Correlation:**</u> • TotalWorkingYears • YearsInCurrentRole • MonthlyIncome
Chi-Squared (X2) Test - 5 Best features correlated to "Income"	Monthly Income TotalWorkingYears YearsAtCompany DailyRate YearsInCurrentRole
Using Feature Importance of Random Forest Classifier (figure 4.7)	JobSatisfaction MonthlyIncome JobInvolvement JobLevel YearsAtCompany

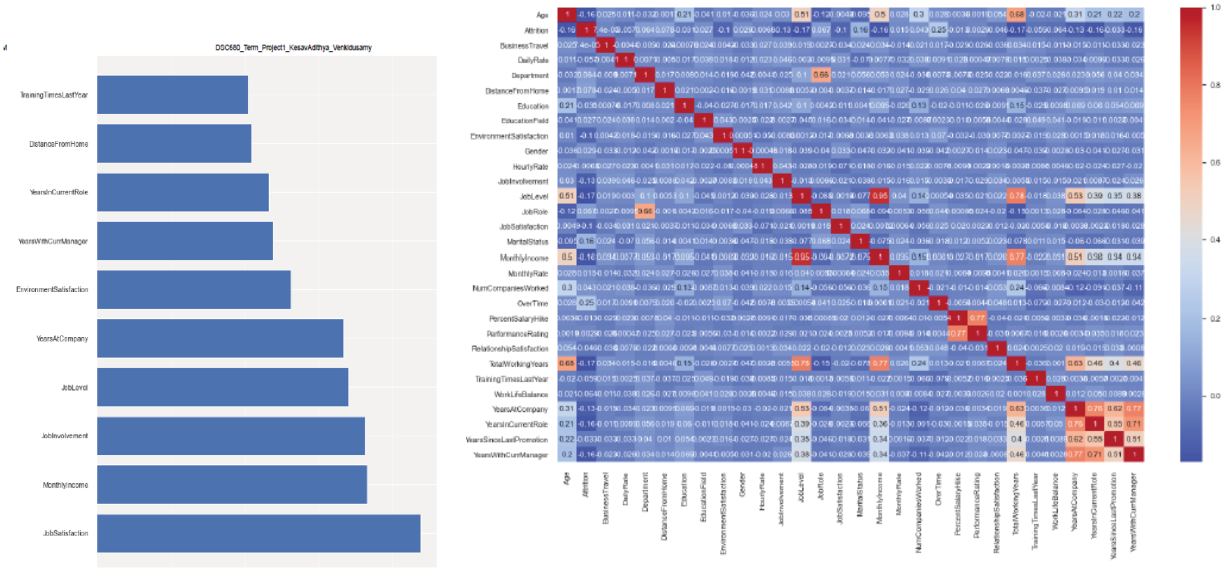


Figure 12 Pearson Correlation Matrix & Random Forest Classifier Feature Significance

Conclusion

Out of three model, Random Forest Classifier

Among various methods used to find 5 best features in the dataset, below are some of the top features having high impact on the target variable "Attrition". So, HR team of an organization can focus on these features to retain their employees.

- Job Satisfaction - Job Satisfaction of the employees
- Monthly Income - Monthly Income earned by the employees
- Years At Company - Employee's experience
- Over time - Over time
- Years in Current Role - Number of years in current role

Assumptions

As part of feature selection, I have assumed the 5 features present in the dataset namely EmployeeNumber, EmployeeCount, StandardHours, Over18, StockOptionLevel doesn't add any value to the target variable "Attrition". So, these features have been removed from the dataset before regression models were built on the data.

Limitations

The dataset considered for this prediction analysis is a fictional dataset created by IBM data scientists. If this dataset doesn't accurately reflect the original real-world data, modeling efforts cannot generate any useful insights.

Challenges

There are couple of challenges I faced during the data preparation step for model building. Identifying the correct features that contribute to the target, planning on how to handle the insufficient data, deciding the next steps if the data is imbalanced to name a few. To mitigate data imbalance, I have chosen "SMOTE" method to over-sample the dataset.

Future Uses/Additional Applications

With the real-world data in similar to this synthetic data, this prediction model can be routinely run to identify employees who are most likely to quit, the key driver of success would be the human element of reaching out the employee, understanding the current situation of the employee and taking action to remedy controllable factors that can prevent attrition of the employee.

Recommendations

Based on the available data, this model predicts the attrition and features impacting the attrition with better accuracy. However, this model should be regressed again when more real-world data is available.

Implementation Plan

With the current features available in the dataset, this model can be implemented to predict the attrition in the organization. In addition, this model can be launched to evaluate various features impacting the attrition. However, as additional features added to the dataset, this model must be reevaluated to ensure there is no slippage due to added features.

Ethical Assessment

One of the ethical considerations for this project is the consideration of results from the analysis in decision-making. Some of the conclusions make from this project's study could be incorrect or misrepresented due to insufficient or incorrect data. So, while sharing the outcome of this project to larger audience, the underlying assumptions and data considerations should be shared.

Another ethical consideration when dealing with employee information is to ensure no personal and sensitive information is present in the dataset. Since this dataset is fictional created by IBM data scientists, this is already taken care by them and personal identifying information (like gender, age) is broad enough which is untraceable to any individual.

References

Pavansubhash (2017). IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Emily Killham (January 25, 2022). Employee Attrition Analytics: The Who, When & Why Of Employee Turnover. https://blog.perceptyx.com/employee-attrition-analytics

Maggie Wooll (January 24, 2022). Fighting employee attrition: What is within your control? https://www.betterup.com/blog/employee-attrition

Unites States Depart of Labor. Annual quits rates by industry and region, not seasonally adjusted. https://www.bls.gov/news.release/jolts.t18.htm