DSCI 5240

DATA MINING

PROJECT- FINAL REPORT

GROUP 11

Names: Jemima Nasara

Sonakshi Sharma

Neelam Gupta

Jouhara Habsi

Ugochi Madumere

Executive Summary

Recruiting, hiring, onboarding, and training new employees' costs IBM billions each year due to employee attrition. They incur losses involving the time, money, and efforts for training in their hiring processes. They also suffer productivity/profit losses when there is constant shed in the workforce, especially top talents and they are very difficult and expensive to replace. Generally, businesses are better off when they can retain good employees and the organizational experience they have. The purpose of this project is to understand what factors affect attrition rates in the employee hiring process and provide valuable insight to IBM's managers in which they can use to better their hiring processes.

We analyzed the "IBM HR Analytics Employee Attrition & Performance"; a fictional dataset from IBM data scientists downloaded from the data source Kaggle. It contained variables that we further analyzed such as "age", "sex", "marital status", "department", "education field", "job level", "job role", "companies worked", "salary hike", "stock option", and "performance rating" that we have understood to be determinants of employee attrition. Our goal is not only reducing the employee attrition rates but providing solutions to management that improve the hiring process.

"Employee Benefit News (EBN) reports that it cost employers 33% of a worker's annual salary to hire a replacement if that worker leaves. In dollar figures, the replacement cost is $15,000 per person for an employee earning a median salary of $45,000 a year, according to the Work Institute's 2017 Retention Report." During 2019, IBM had 352,600 employees. Based on the analysis about the "IBM HR Analytic Employee Attrition & Performance", we discovered some valuable insights, most of the employees who left the company have worked for 0-5 years with relatively low monthly income. Although some employees seem satisfied with the work environment, colleagues, and culture, they still choose to leave. Another interesting finding about the company's employees is that age and work experience are highly correlated with income in the company. This partly explains why employees leave. Most people who left have worked for a short period of time, which relates to lower salary.

The focus of our project is aimed at helping IBM human resources make better hiring decisions and find resignation trends/numbers to understand the latent factors which contribute to employee attrition in the organization. We have used different models to predict which variables affect employee attrition rates. We have used the DMAIC problem solving method to define, measure, analyze, improve, and control IBMs hiring process. Our study provides insight to reduce financial losses due to employee attrition as we assist with proactiveness in employee welfare planning.

Dataset Description

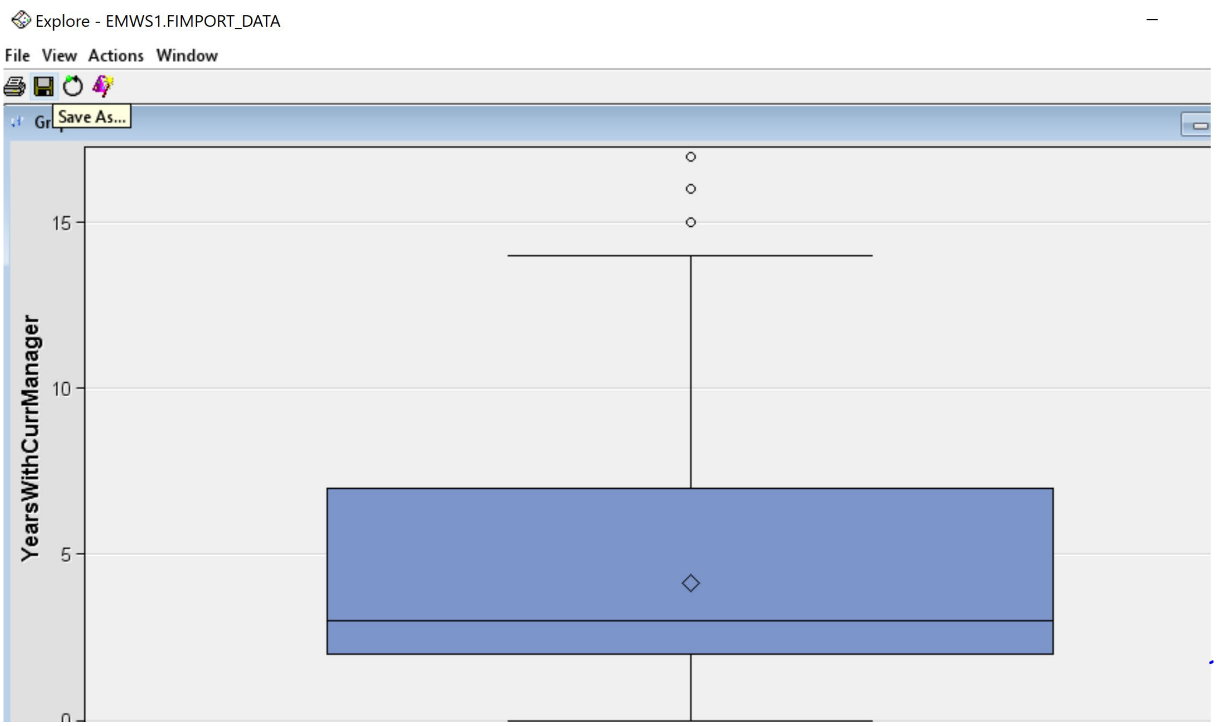
We used a dataset called "IBM HR Analytics Employee Attrition & Performance"; a fictional dataset from IBM data scientists downloaded from the data source Kaggle. This dataset is composed of 1479 variables and 35 columns and is available for download at https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/data#. It took about two days to find this dataset after a rigorous search. This data has been preprocessed, so there were no missing data found when we attempted to clean our dataset.

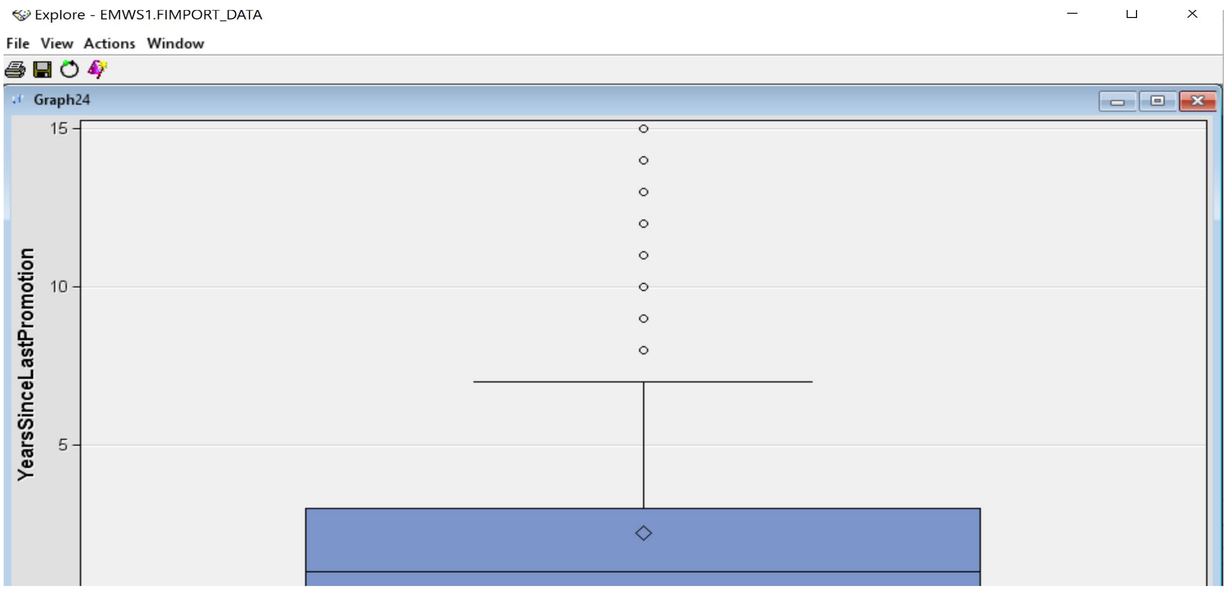
Dataset Preparation

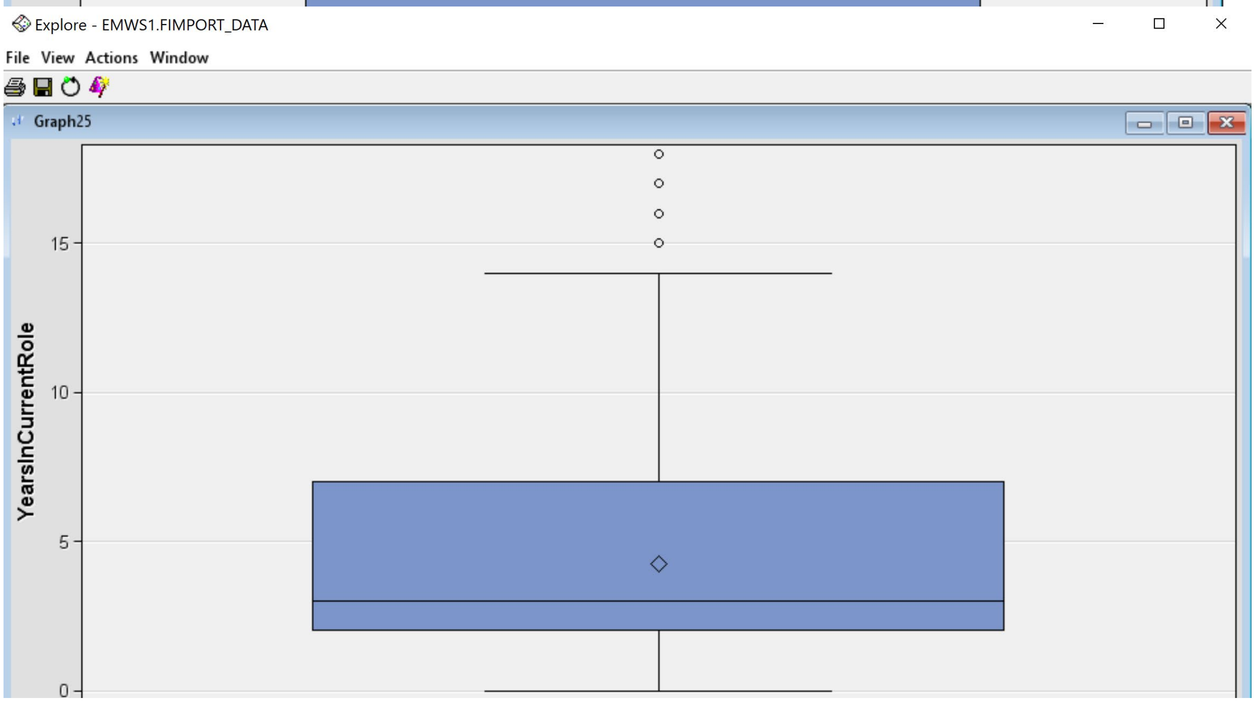
The following variables were rejected: Over 18 and Employee Count as they contributed no significant information to the dataset due to common values across the features. Outliers were detected in the following features Years since Last Promotion, Years with Current Manager, Years in Current Role, Years at Company, Stock Option Level, Training Times Last Year, Number of Companies Worked, Total Working Years, and Monthly Income. We applied a Filter node to remove the outliers before commencing the modeling steps. Dataset was split into 60% train, 20% test and 20% validation prior to modeling using the data partition node.

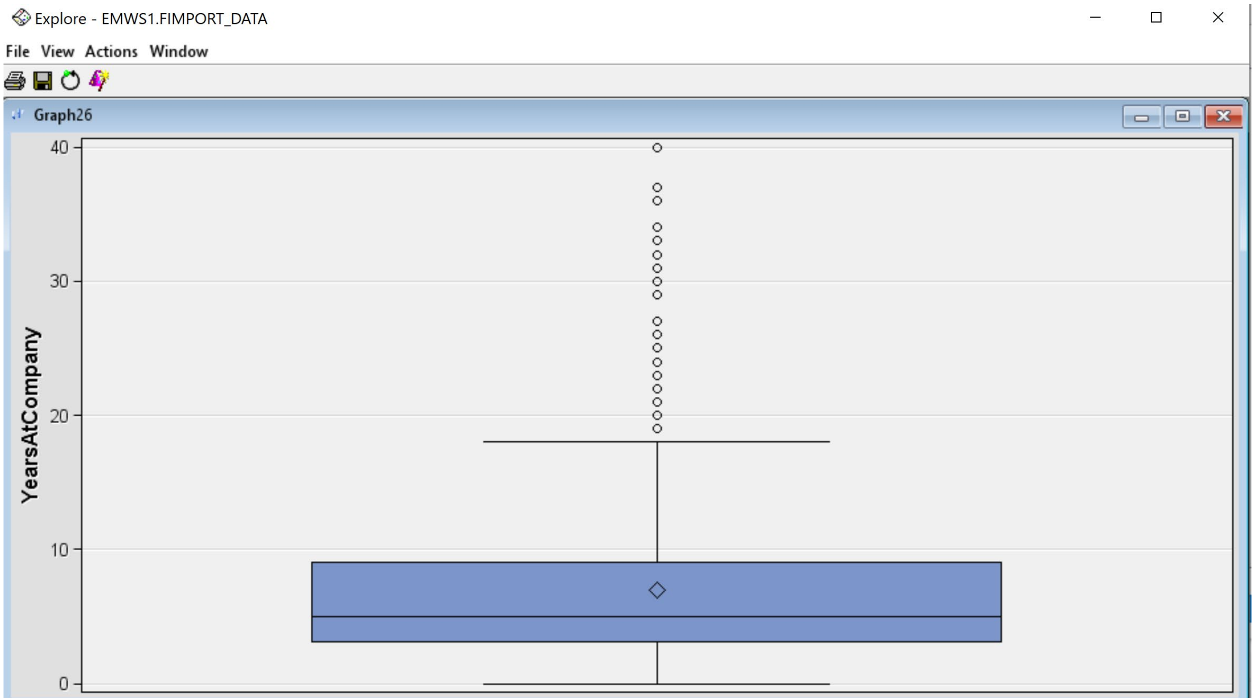
Exploratory Data Analysis

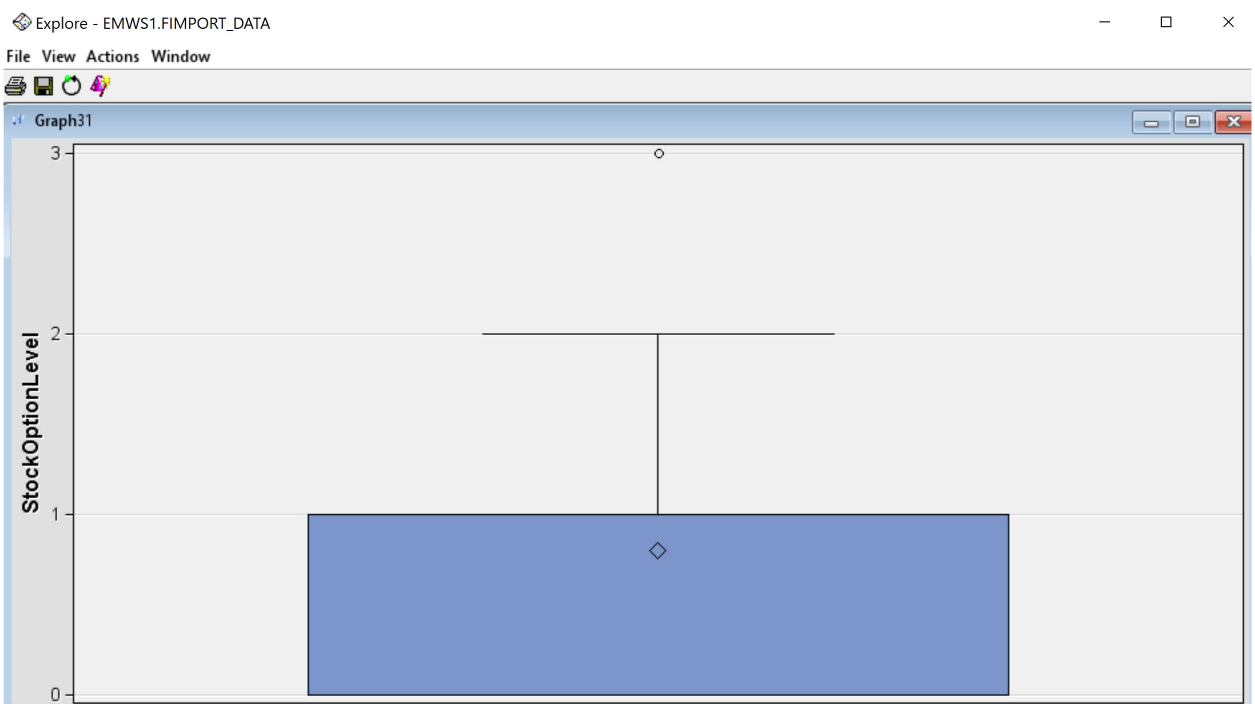
Outlier Detection

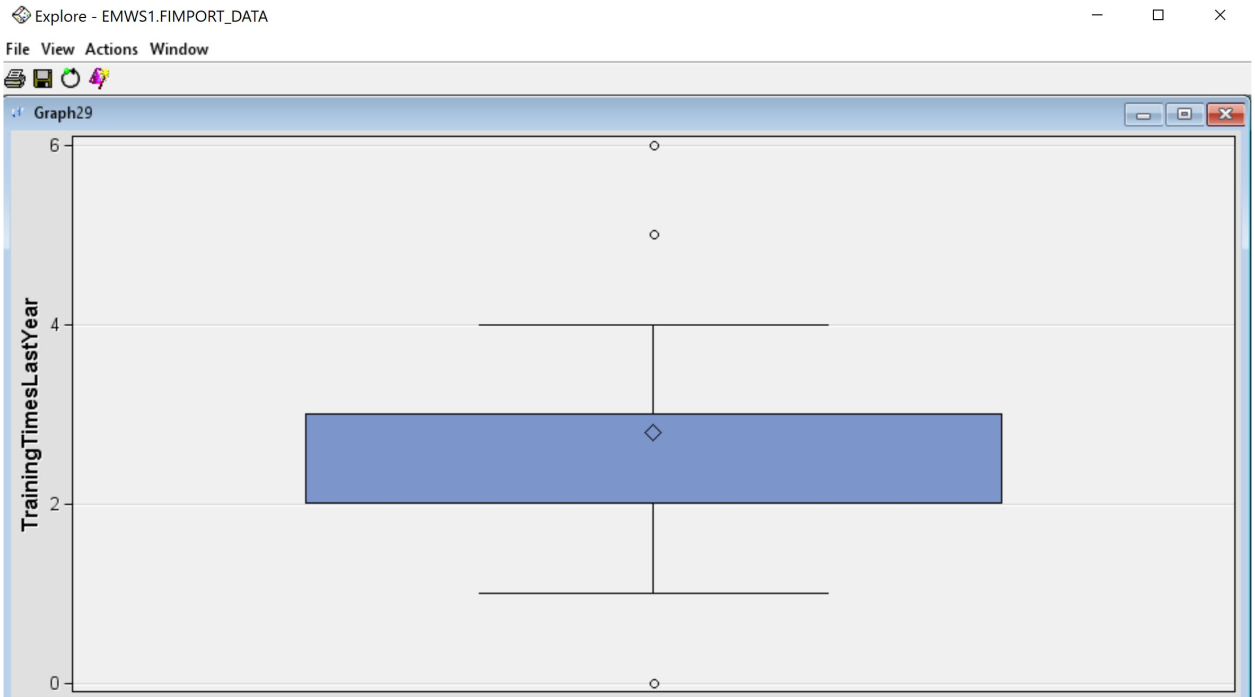


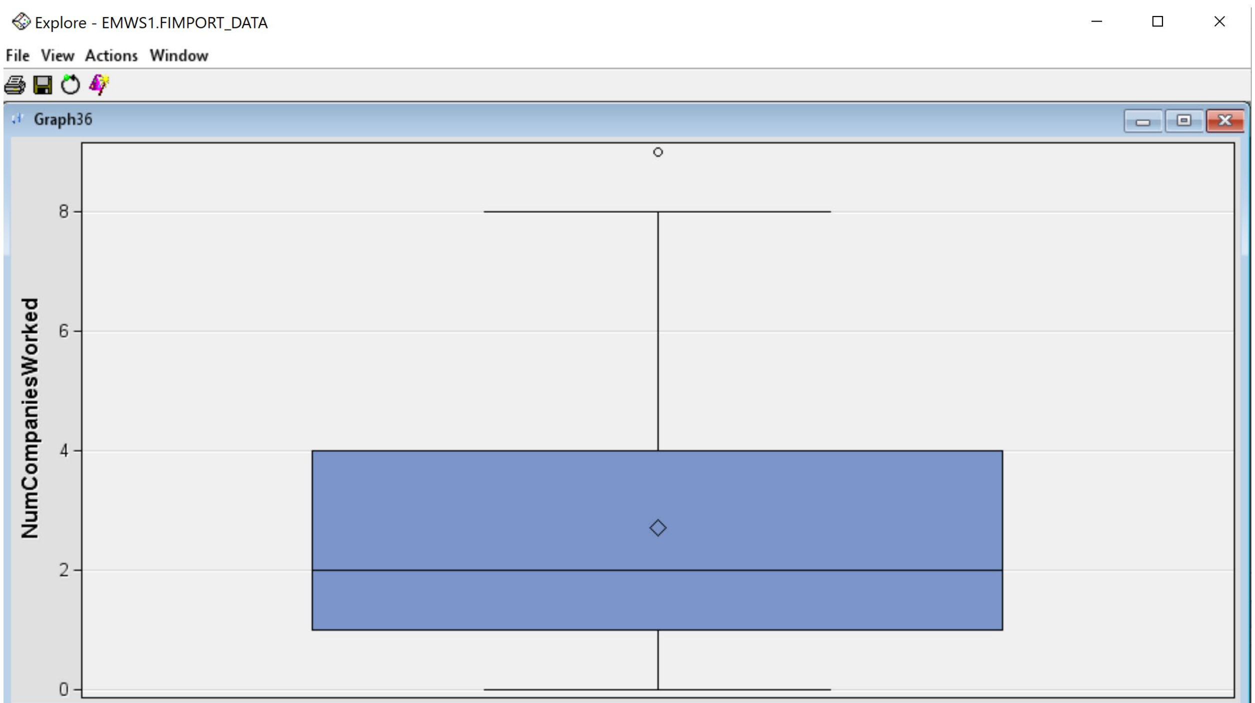

Explore - EMWS1.FIMPORT_DATA
File View Actions Window
Graph24
YearsSinceLastPromotion
15
10
5

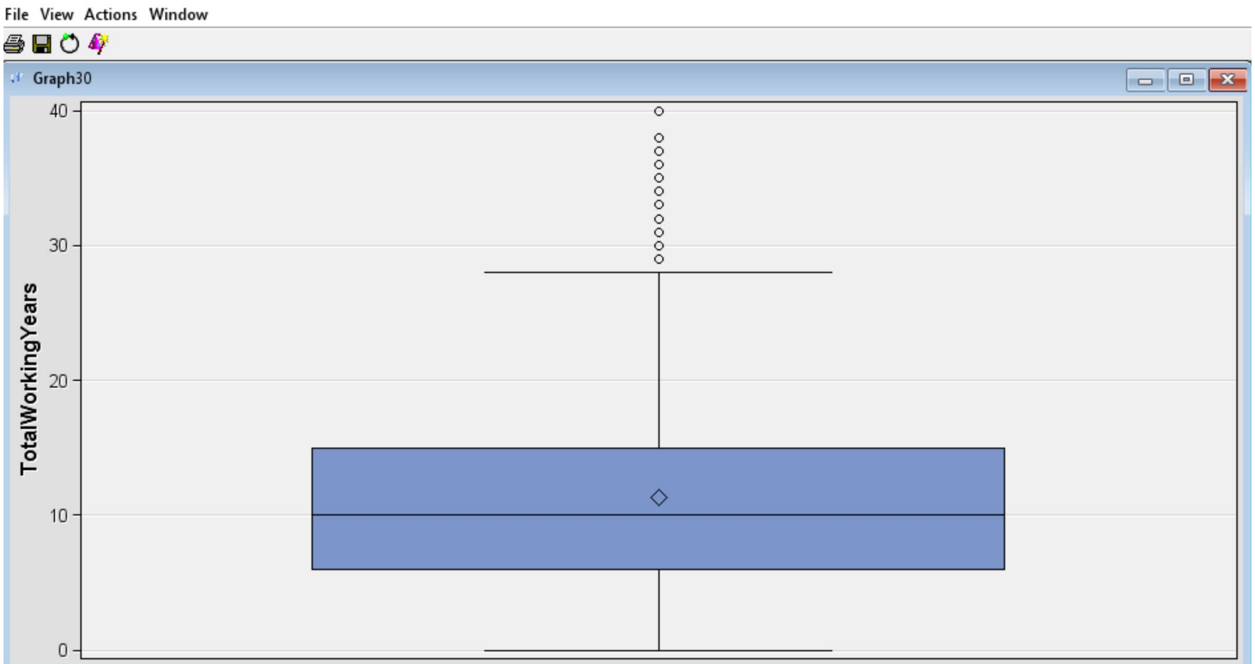

Explore - EMWS1.FIMPORT_DATA
File View Actions Window
Graph25
YearsInCurrentRole
15
10
5
0


Explore - EMWS1.FIMPORT_DATA
File View Actions Window
Graph26
YearsAtCompany
40
30
20
10
0


Explore - EMWS1.FIMPORT_DATA
File View Actions Window
Graph31
StockOptionLevel
3
2
1
0


Explore - EMWS1.FIMPORT_DATA
File View Actions Window
Graph29
TrainingTimesLastYear
6
4
2
0


Explore - EMWS1.FIMPORT_DATA
File View Actions Window
Graph36
NumCompaniesWorked
8
6
4
2
0



We performed exploratory analysis between different attributes against our target variable (Attrition rate) before creating a model. This will help us to understand our data in detail and create a better model.

These were some of the questions we asked ourselves?

1) What is the job satisfaction by attrition rate?

2) What is the distribution of gender against attrition rate?

3) Does age play a major role in the Attrition rate?

4) Does the average hourly rate play a major role in the Attrition rate?

5) Does frequent work-related travel play a major role in the average hourly rate?

6) Does education play a major role in the attrition rate?

7) Does Job involvement play a major role in the attrition rate?

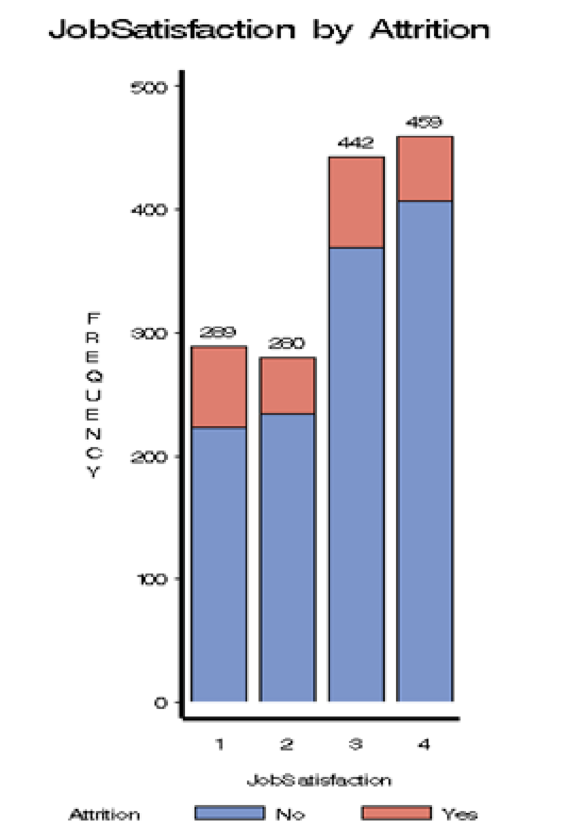
8) Does job satisfaction play a major role in the attrition rate?

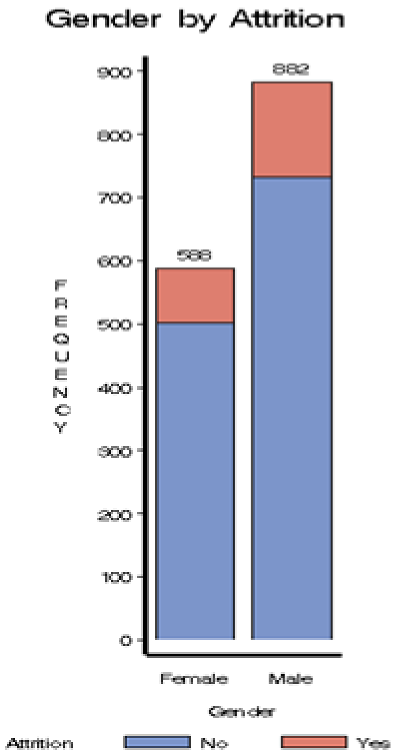
9) Does overtime have any role in attrition rate?

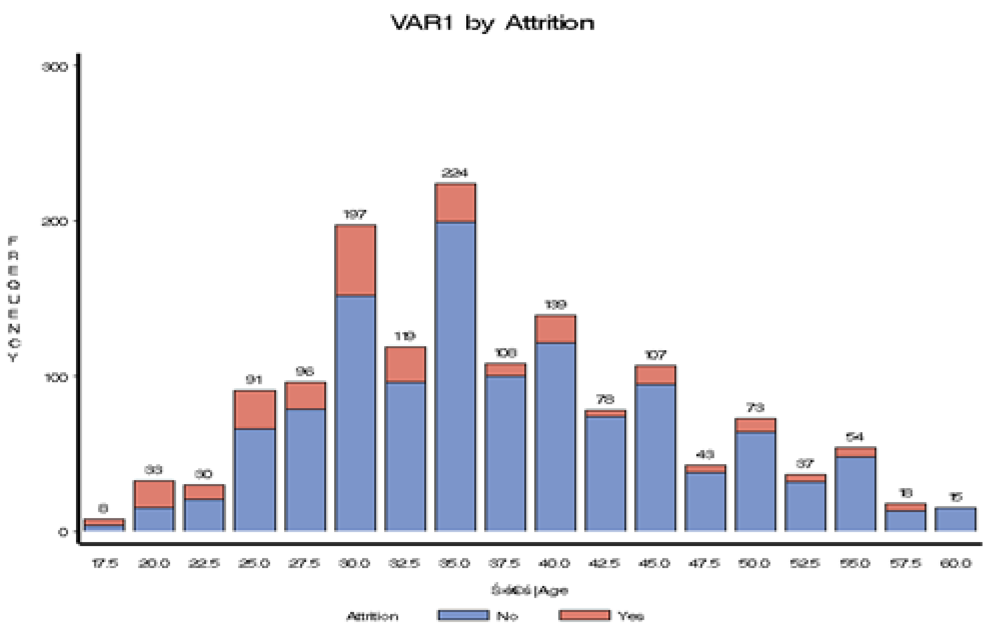
10) Is performance rating plays a major role in the attrition rate?

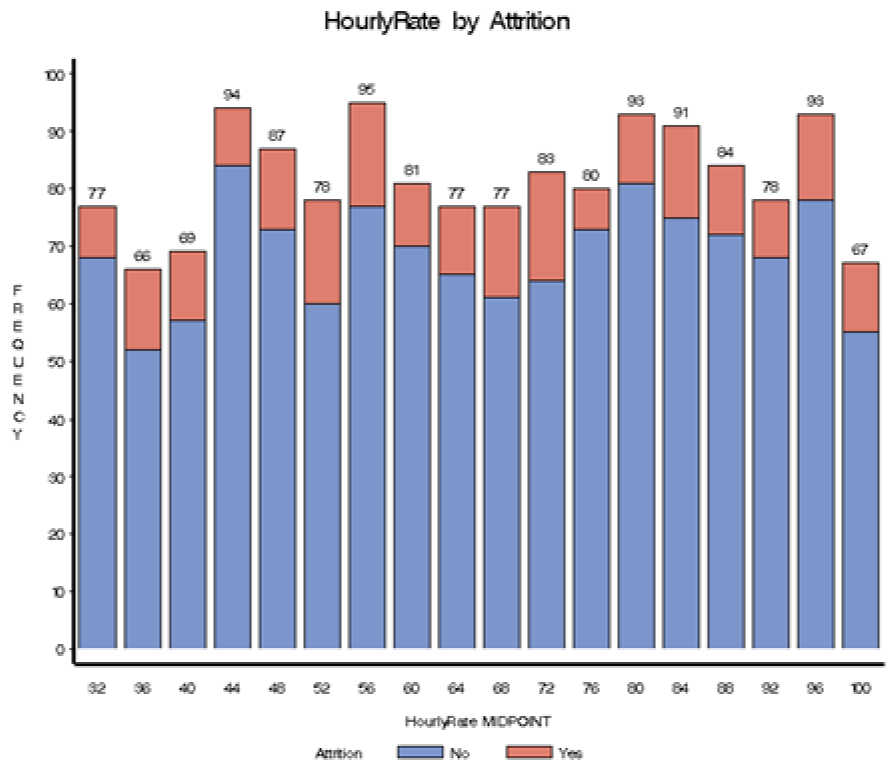
11) Does the number of years working at a company play a role in the attrition rate?

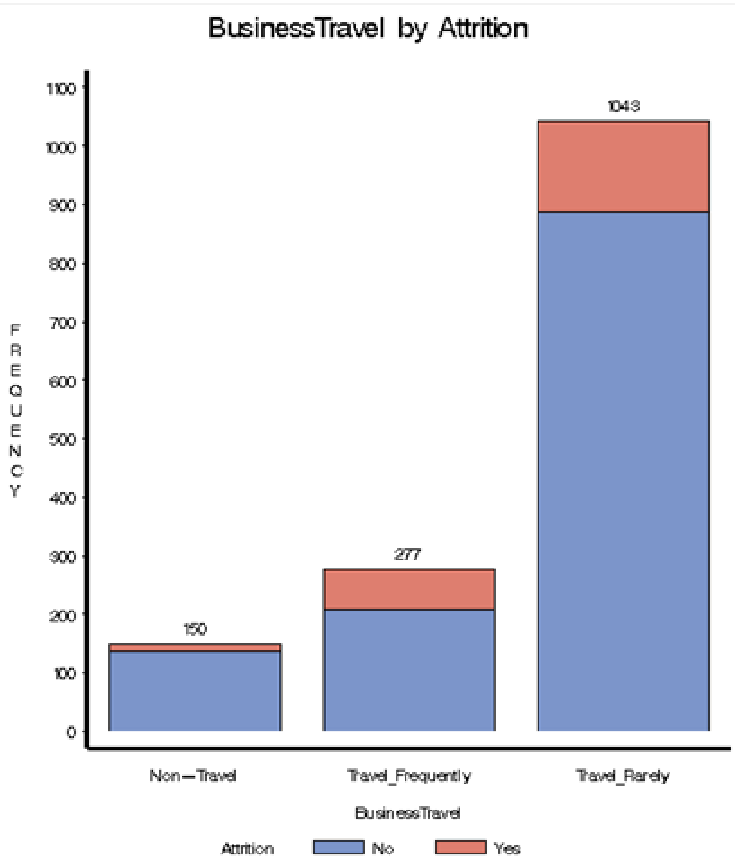
Bar Charts

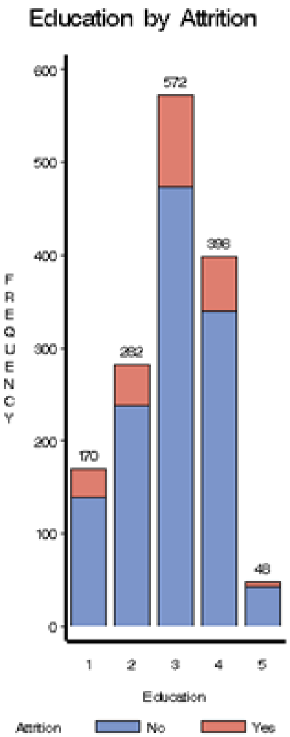


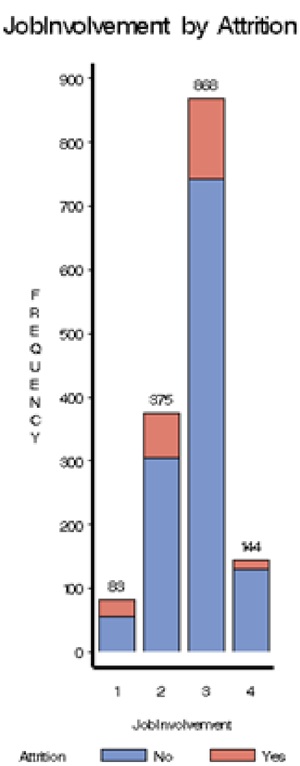

Gender by Attrition
FREQUENCY
0
100
200
300
400
500
600
700
800
900
588
882
Female
Male
Gender
Attrition
No
Yes

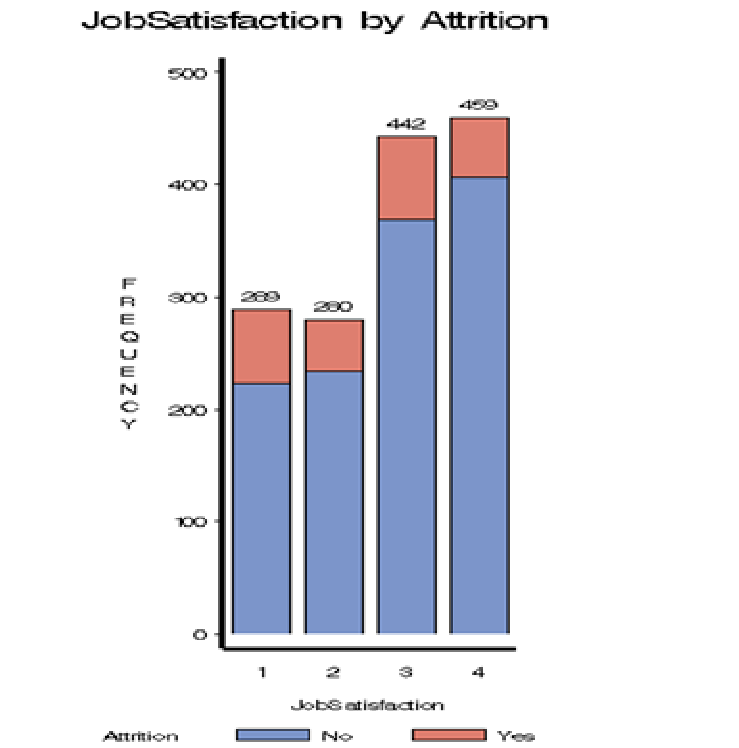

VAR1 by Attrition
FREQUENCY
0
100
200
300
8
33
30
91
96
197
119
224
108
139
78
107
43
73
37
54
18
15
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0
Age
Attrition
No
Yes

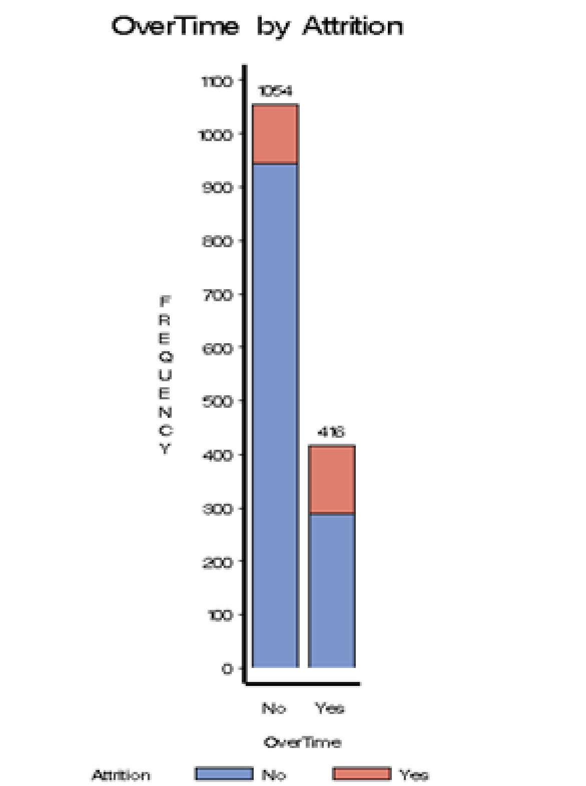

HourlyRate by Attrition
FREQUENCY
0
10
20
30
40
50
60
70
80
90
100
77
66
69
94
87
78
95
81
77
77
83
80
93
91
84
78
93
67
32
36
40
44
48
52
56
60
64
68
72
76
80
84
88
92
96
100
HourlyRate MIDPOINT
Attrition
No
Yes

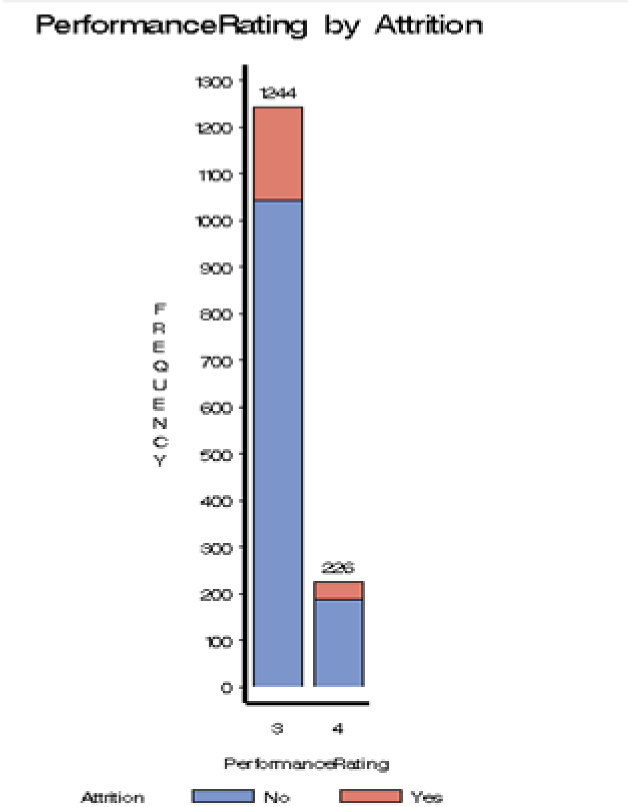

BusinessTravel by Attrition
FREQUENCY
0
100
200
300
400
500
600
700
800
900
1000
1100
150
277
1043
Non-Travel
Travel_Frequently
Travel_Rarely
BusinessTravel
Attrition
No
Yes

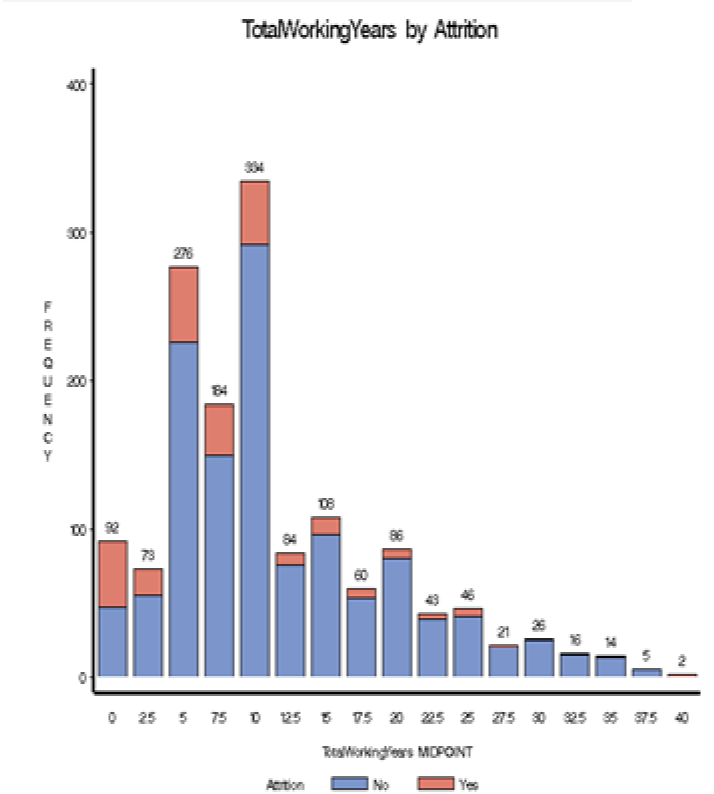

Education by Attrition
FREQUENCY
600
500
400
300
200
100
0
572
398
282
170
48
1
2
3
4
5
Education
Attrition
No
Yes


JobInvolvement by Attrition
FREQUENCY
900
800
700
600
500
400
300
200
100
0
868
375
144
83
1
2
3
4
JobInvolvement
Attrition
No
Yes


JobSatisfaction by Attrition
FREQUENCY
500
400
300
200
100
0
289
280
442
459
1
2
3
4
JobSatisfaction
Attrition
No
Yes


OverTime by Attrition
FREQUENCY
1100
1000
900
800
700
600
500
400
300
200
100
0
1054
416
No
Yes
OverTime
Attrition
No
Yes



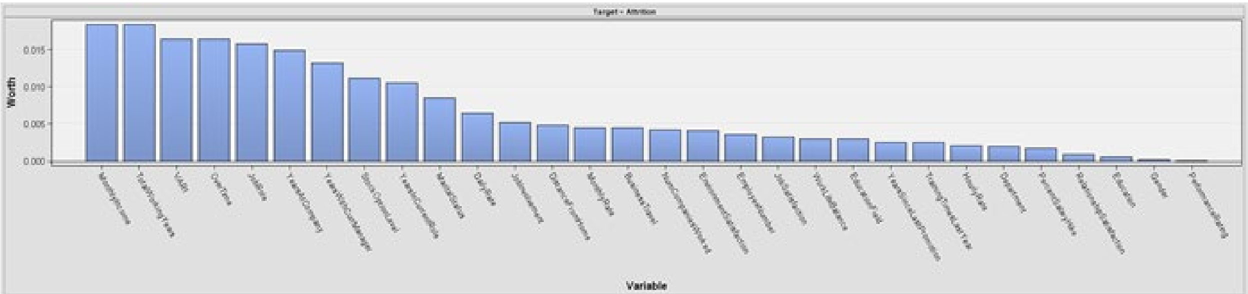


Summary:

- Attrition rate is higher in males than females

- Attrition rate is higher in the age group between 30 to 35 followed by the age of 40.0 It shows us that people within this age group switch jobs due to salary or job satisfaction. The higher the age group, the lesser the attrition rate.
- There is no pattern in the average hourly rate against the attrition rate. Hence, we can't say that hourly rate has a significant role in the attrition rate.
- Attrition rate is low when employees travel frequently. However, attrition rate is high when people travel rarely. This may be a reason these employees might be in the administration or different department where job roles don't require any frequent travel. We cannot conclude that frequent travel plays any major role in the attrition rate.
- Employees who have higher education have a lower attrition rate and employees who have a mid-level education have a higher attrition rate.
- Education (1 -> Below College, 2 -> College, 3 -> Bachelor, 4 -> Master 5 -> Doctor). Bachelor and Master's degree employees have a high attrition rate. However, attrition rate is less for doctors and below college graduates' employees. It is clear that bachelor and master's degree employees are getting jobs and easily switching their jobs.
- Job Involvement (1 -> Low, 2 -> Medium, 3 -> High, 4 -> Very High). High job involvement employees have a high attrition rate.
- Job Satisfaction (1 -> Low, 2 -> Medium, 3 -> High, 4 -> Very High). High and very high job satisfaction employees have a high attrition rate. We can say that even though these employees have high job involvement, they are leaving the company, maybe because they are not faced with enough challenges in the company, or they are getting higher positions from the other companies.
- Overtime does not play any major role in the attrition rate.
- It is clear from the graph that Excellent employees (3 -> Excellent, 4 -> Outstanding) have a high attrition rate.
- Employees who work 5 to 10 years in the company have a high attrition rate.

Variable worth



The variable worth plot orders the variables by their worth in predicting the target variable based on the Gini Split worth statistic. The plot shows Monthly income, Total working.

Descriptive Statistics

Interval Variables

Obs	NAME	NMISS	N	MIN	MAX	MEAN	STD	SKEWNESS	KURTOSIS
1	DailyRate	0	1470	102	1499	802.49	403.51	-0.00352	-1.20382
2	DistanceFromHome	0	1470	1	29	9.19	8.11	0.95812	-0.22483
3	Education	0	1470	1	5	2.91	1.02	-0.28968	-0.55911
4	EmployeeNumber	0	1470	1	2068	1024.87	602.02	0.01657	-1.22318
5	EnvironmentSatisfaction	0	1470	1	4	2.72	1.09	-0.32165	-1.20252
6	HourlyRate	0	1470	30	100	65.89	20.33	-0.03231	-1.19640
7	JobInvolvement	0	1470	1	4	2.73	0.71	-0.49842	0.27100
8	JobLevel	0	1470	1	5	2.06	1.11	1.02540	0.39915
9	JobSatisfaction	0	1470	1	4	2.73	1.10	-0.32967	-1.22219
10	MonthlyIncome	0	1470	1009	19999	6502.93	4707.96	1.36982	1.00523
11	MonthlyRate	0	1470	2094	26999	14313.10	7117.79	0.01858	-1.21496
12	NumCompaniesWorked	0	1470	0	9	2.69	2.50	1.02647	0.01021
13	PercentSalaryHike	0	1470	11	25	15.21	3.66	0.82113	-0.30060
14	PerformanceRating	0	1470	3	4	3.15	0.36	1.92188	1.69594
15	RelationshipSatisfaction	0	1470	1	4	2.71	1.08	-0.30283	-1.18481
16	StockOptionLevel	0	1470	0	3	0.79	0.85	0.96898	0.36463
17	TotalWorkingYears	0	1470	0	40	11.28	7.78	1.11717	0.91827
18	TrainingTimesLastYear	0	1470	0	6	2.80	1.29	0.55312	0.49499
19	VAR1	0	1470	18	60	36.92	9.14	0.41329	-0.40415
20	WorkLifeBalance	0	1470	1	4	2.76	0.71	-0.55248	0.41946
21	YearsAtCompany	0	1470	0	40	7.01	6.13	1.76453	3.93551
22	YearsInCurrentRole	0	1470	0	18	4.23	3.62	0.91736	0.47742
23	YearsSinceLastPromotion	0	1470	0	15	2.19	3.22	1.98429	3.61267
24	YearsWithCurrManager	0	1470	0	17	4.12	3.57	0.83345	0.17106

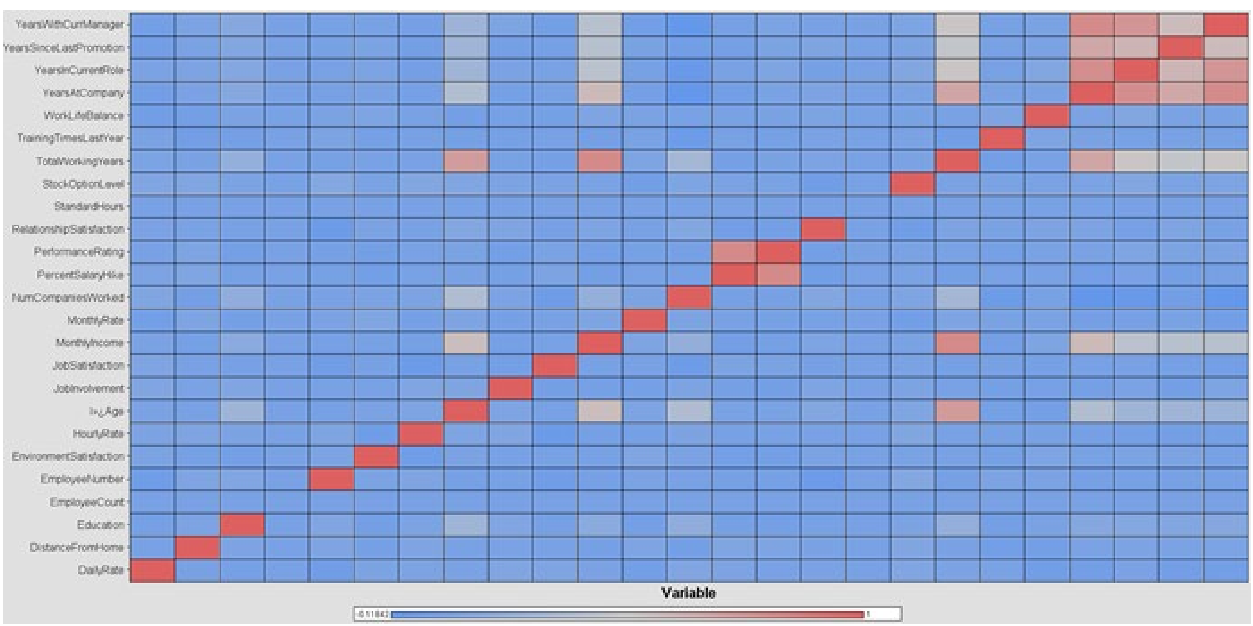
Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Attrition	INPUT	2	0	No	83.88	Yes	16.12
TRAIN	BusinessTravel	INPUT	3	0	Travel_Rarely	70.95	Travel_Frequently	18.84
TRAIN	Department	INPUT	3	0	Research & Development	65.37	Sales	30.34
TRAIN	EducationField	INPUT	6	0	Life Sciences	41.22	Medical	31.56
TRAIN	JobRole	INPUT	9	0	Sales Executive	22.18	Research Scientist	19.86
TRAIN	MaritalStatus	INPUT	3	0	Married	45.78	Single	31.97
TRAIN	OverTime	INPUT	2	0	No	71.70	Yes	28.30

The above descriptive statistics shows that there are no missing values and the data is not normally distributed as the skewness and kurtosis values are not in the rage of +/- 1.96. Our goal is to apply the Decision Tree algorithm to this problem, for any tree-based algorithm, we can proceed building the model without normalization of the data. The main reason is that a tree-based model makes decisions at a node based on a single feature at a time, hence difference scales of features and outliers won't impact the algorithm.

Correlation Matrix

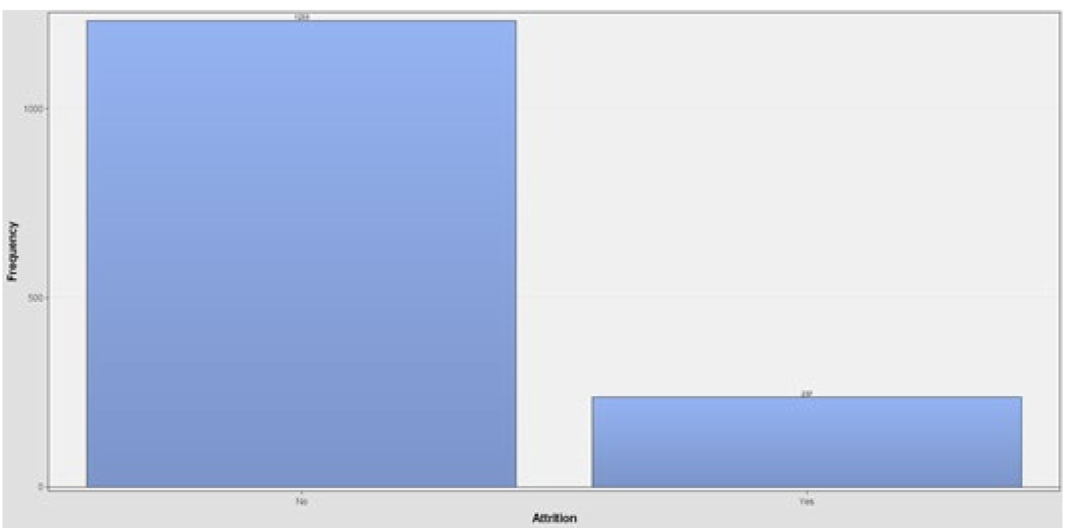


Generated above is correlation matrix from the variable clustering node. Below is the analysis:

- The higher the monthly income, the higher the total working years of an employee.
- The higher the performance rating, the higher the percent salary hike.
- The higher the years since the last promotion, the higher the years with the current manager.
- The higher the monthly income, the higher the age.
- The higher the current company, the higher in the current role

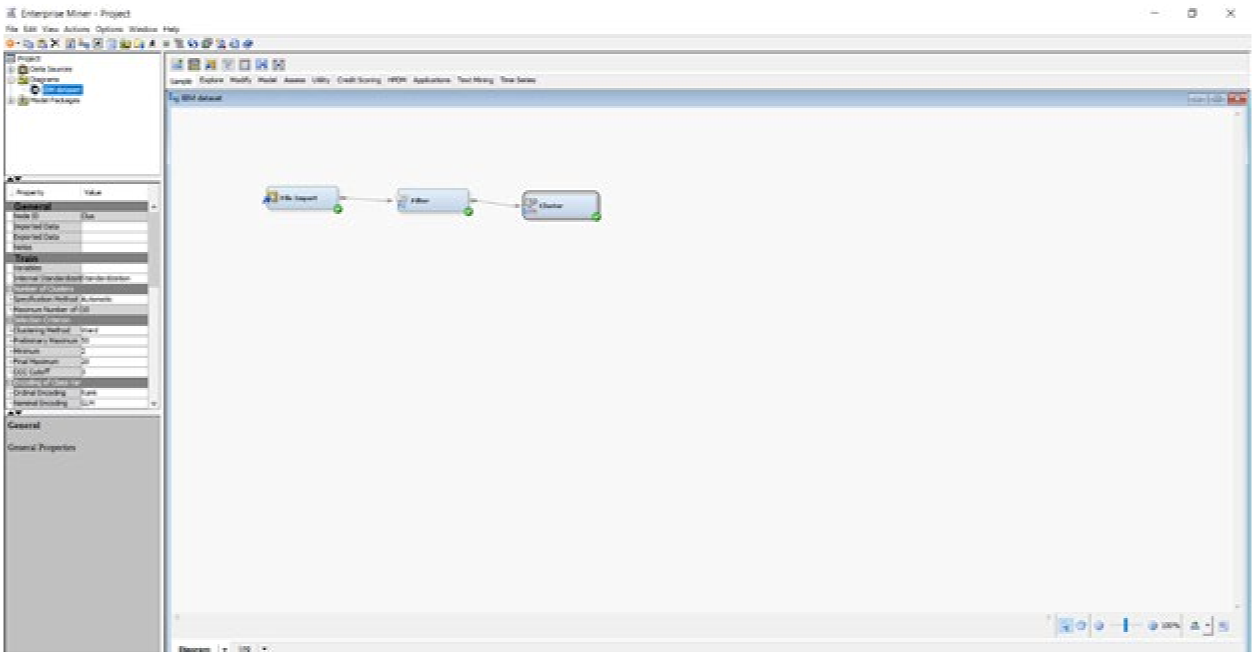
After removing total working years, years with current manager, years in current role from our model, it is highly correlated with current company, higher the years and monthly income.

Target Variable (y)



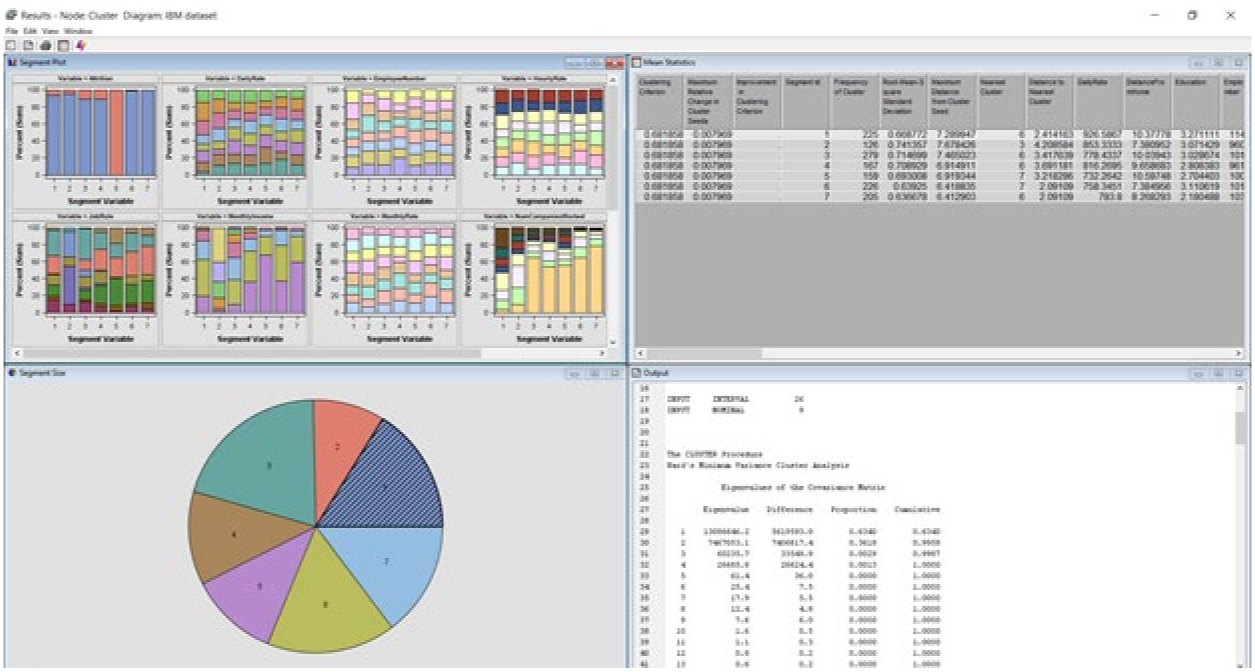
The target variable is Attrition Rate. It is defined by a classification of attrition rate with two classes, ‘yes’ and ‘no’. The above figure shows that 80% ‘No’ and 20% ‘Yes’, which clearly shows that our dataset is heavily imbalanced, and it does not represent the attrition rate equally. This may result in model accuracy being incorrectly measured, depending on the metric used. There was no resampling used and AUC (Area Under Curve) and misclassification rate was used as an assessment parameter.

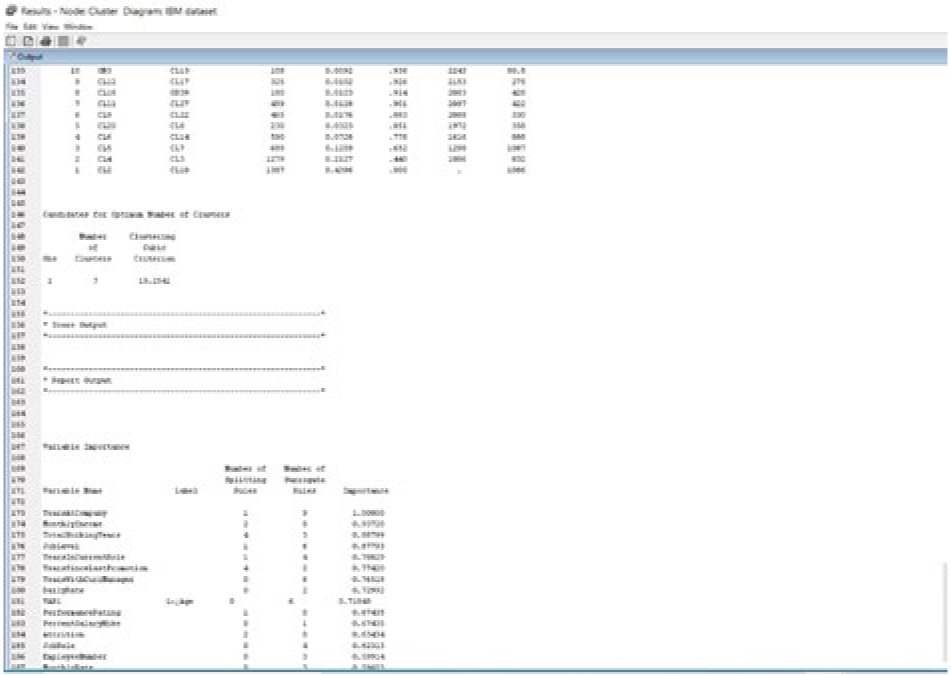
Cluster Analysis



Here, the filter node is used to exclude certain observations, such as extreme outliers. This is because filtering extreme values from the training data tends to produce better models because the parameter estimates will be more stable.

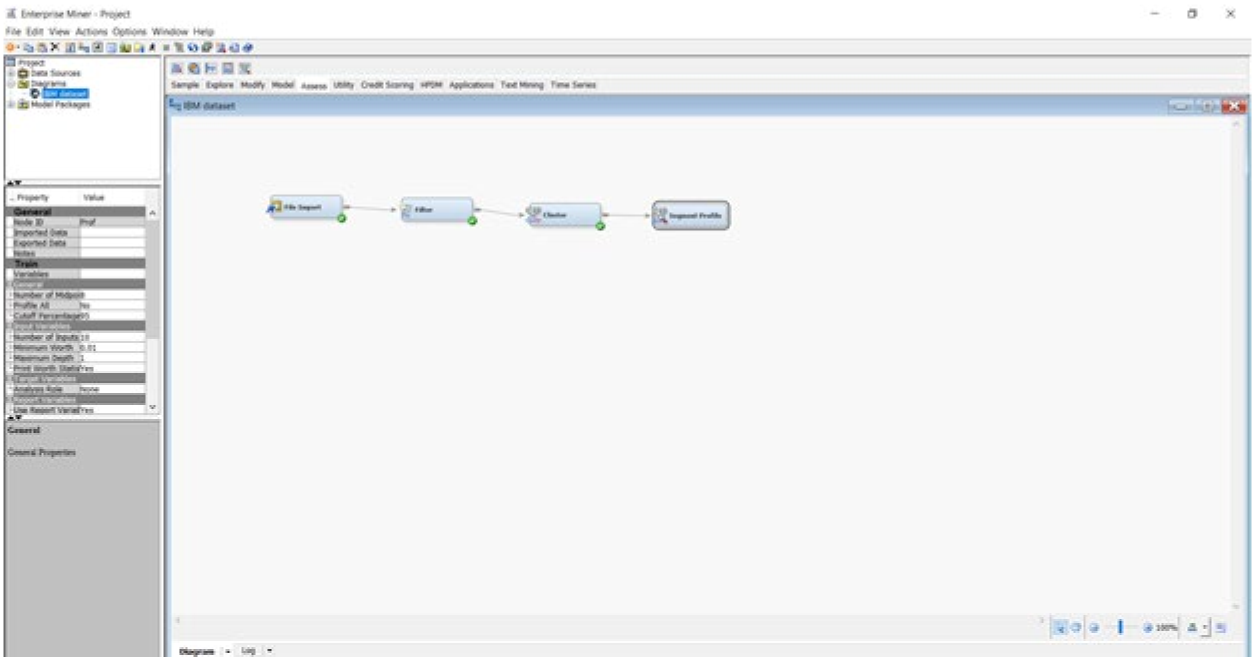
Results:

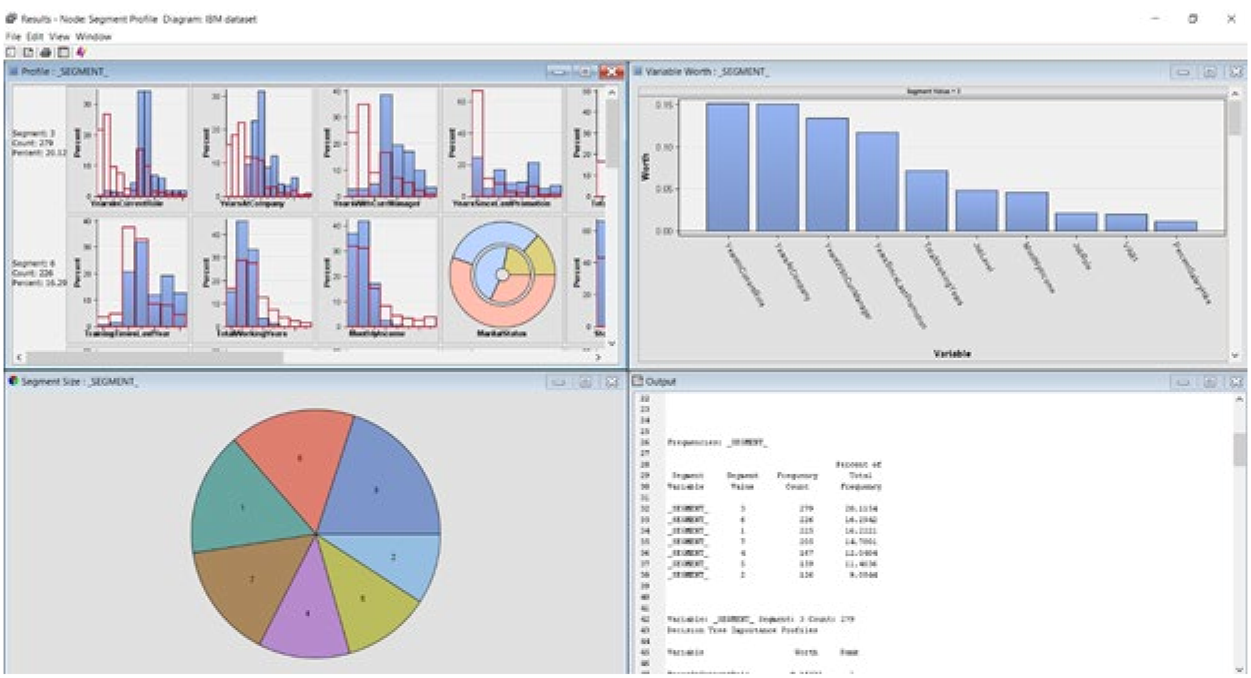


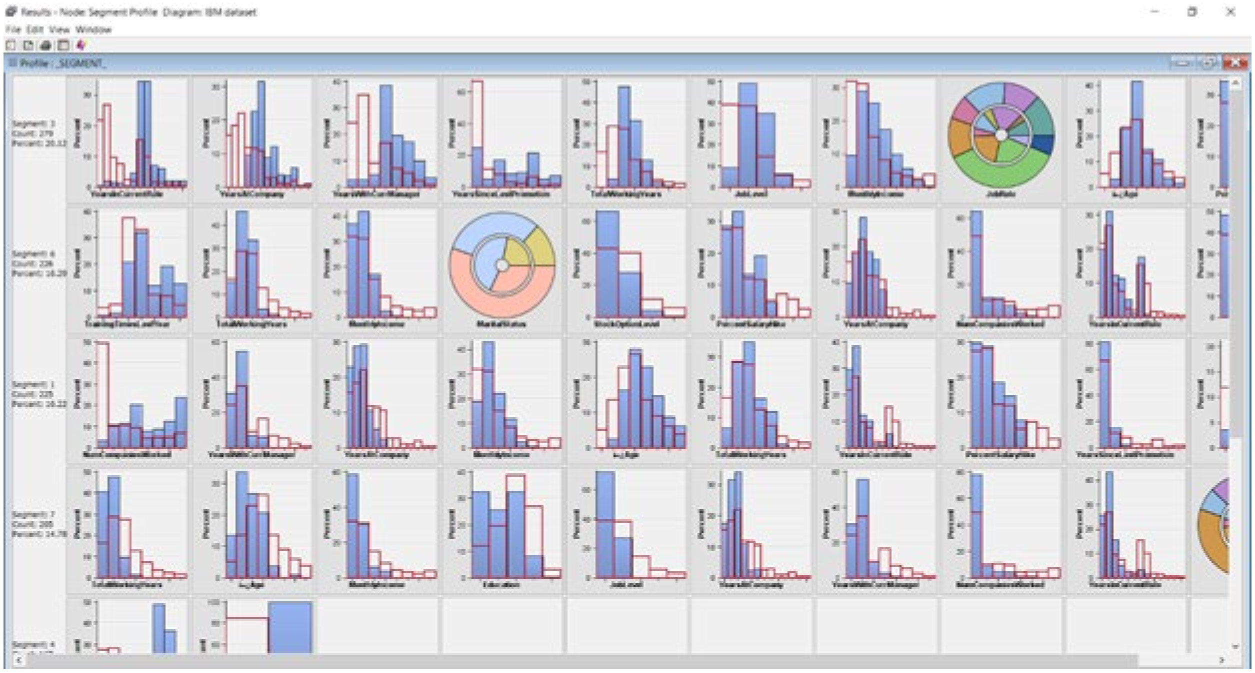


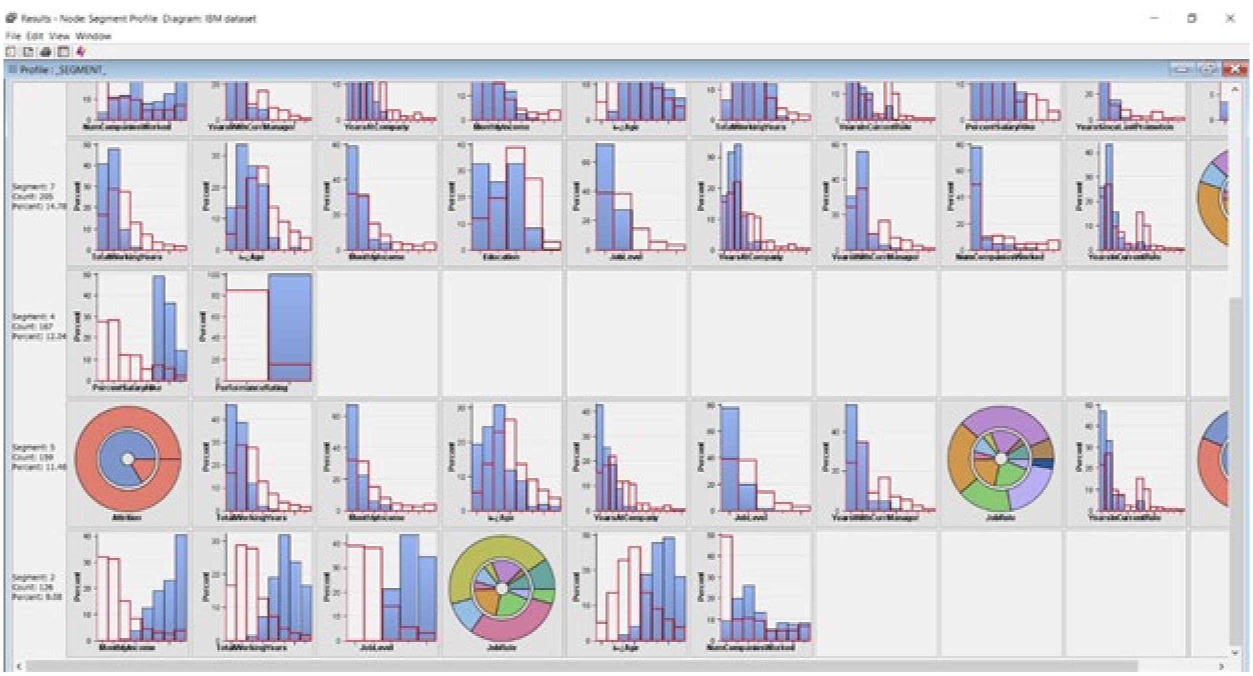
After running the Cluster Node, it has a result of 7 clusters and frequency ranging from 126 to 279. The frequency of each cluster is quite reasonable because we have a uniform spread of observations across clusters, except for clusters 2, 4 and 5 which have 126, 167 and 159 observations, while every other cluster is within the range of 200-279. For example, the frequency of segment 1 is 225, segment 3 has a frequency of 279, segment 6 is 226, and segment 7 has a frequency of 205. Measuring the standard deviations of all seven clusters varies between 0.6 and 0.7. Having all the values fit between 7 clusters is also an indicator of high intra cluster similarity and low inter cluster similarity. The clusters seem meaningful, and all are sized well.

After running the Segment Profile Node:









Segment 1

Cluster 1 is responsible for 16.22% of the entire sample.

The different variables for this segment have been allocated a worth based on their presence in the cluster. The sequence from highest worth to lowest worth in segment 1 is

Number of Companies worked, Years with Current Manager, Years at Company, Monthly Income, Age, Total Working Years, Years in Current Role, Percent Salary Hike, Years Since Last Promotion, and Daily Rate.

The histograms indicate observations in the cluster, as well as the observations present in the dataset. Looking at the histograms, it indicates that from the observations in the cluster, the mean for the variable Number of Companies worked is lower than the mean of observations in the dataset. On the other hand, the mean for the observations in the cluster Age is the same as the mean for the observations present in the dataset.

This indicates that the variable Number of Companies Worked is the highest contributor of HR attrition in a company.

Segment 2

Cluster 2 is responsible for 9.08% of the entire sample and therefore records the lowest percentage compared to the other clusters.

The different variables for this segment have been allocated a worth based on their presence in the cluster. The sequence from highest to lowest worth in segment 2 is Monthly Income, Total Working Years, Job Level, Job Role, Age and Number of Companies Worked.

The histograms indicate observations in the cluster, as well as the observations present in the dataset. Looking at the histograms, this indicates that from the observations in the cluster, the mean for the variable Monthly Income is higher than the mean of observations in the dataset. On the other hand, the mean in the Number of Companies worked would be almost the same as the mean present in the dataset because of the level of skewness. The histogram for the cluster is right skewed which makes the mean closer to the tail.

This indicates that the cluster Monthly Income would be another contributor of attrition in a company.

Segment 3

Cluster 3 is responsible for 20.12% of the entire sample and therefore records the highest percentage compared to other clusters.

The different variables for this segment have been allocated a worth based on their presence in the cluster. The sequence from highest worth to lowest worth in segment 3 is Years in Current Role, Years at Company, Years with Current Manager, Years since Last Promotion, Total Working Years, Job level, Monthly Income, Job Role, Age, and Percent Salary Hike.

The histograms indicate observations in the cluster, as well as the observations present in the dataset. For variables (Monthly Income, Age and Percent Salary Hike), the mean for the observations in the dataset for the variables is like the mean of observations of these variables present in the cluster.

This indicates that the clusters will also be contributors of HR attrition in a company.

Segment 4

Cluster 4 is responsible for 12.04% of the entire sample.

The different variables for this segment have been allocated a worth based on their presence in the cluster. The sequence from highest worth to lowest worth in segment 4 is Percent Salary Hike and Performance Rating.

The histograms indicate observations in the cluster, as well as the observations present in the dataset. Looking at the histograms, we can see that the mean for the observations in the dataset for the variable Percent Salary Hike is higher than the observations of these variables present in the cluster. For another variable (Performance Rating), the mean for the observations in the dataset is lower than the observations of these variables present in the cluster.

This indicates that the clusters Percent Salary Hike and Performance Rating are both contributors to attrition in a company.

Segment 5

Cluster 5 is responsible for 11.46% of the entire sample.

The different variables for this segment have been allocated a worth based on their presence in the cluster. The sequence from highest worth to lowest worth in segment 5 is Attrition, Total Working Years, Monthly Income, Age, Years at Company, Job Level, Years with Current Manager, Job Role, Years in Current Role, and Overtime.

The histograms indicate observations in the cluster, as well as the observations present in the dataset. Looking at the histograms, we can see that the mean for the observations in the dataset for the variables Total Working Years and Monthly Income are lower than the observations of these variables present in the cluster. On the other hand, the mean for the observations in the dataset for the variable Age would be similar to the mean of the observations in the cluster.

This indicates that most of the clusters for example Total Working Years, Monthly Income and Years at Company would be some of the lowest contributing factors to HR attrition in a company.

Segment 6

Cluster 5 is responsible for 16.29% of the entire sample.

The different variables for this segment have been allocated a worth based on their presence in the cluster. The sequence from highest worth to lowest worth in segment 6 is Training Times Last Year, Total Working Years, Monthly Income, Marital Status, Stock Option Level, Percent Salary Hike, Years at Company, Number of Companies Worked, Years in Current Role, Job Level.

The histograms indicate observations in the cluster, as well as the observations present in the dataset. Looking at the histograms, we can see that the mean for the observations in the dataset for the variables Monthly Income and Stock Option level are lower than the observations of these variables present in the cluster. On the other hand, the mean for the observations in the dataset for the variable Years in Current Role would be similar to the mean of the observations in the cluster.

This indicates that most of the clusters for example Monthly Income and Stock Option would be some of the lowest contributing factors to HR attrition in a company.

Segment 7

Cluster 7 is responsible for 14.78% of the entire sample and therefore records the highest percentage compared to other clusters.

The different variables for this segment have been allocated a worth based on their presence in the cluster. The sequence from highest worth to lowest worth in segment 7 is Total Working Years, Age, Monthly Income, Education, Job Level, Years at Company, Years with Current Manager, Number of Companies Worked, Years in Current Role, and Job Role.

The histograms indicate observations in the cluster, as well as the observations present in the dataset. For variables (Number of Companies Worked and Job Level), the mean for the observations in the dataset for the variables is lower than the mean of observations of these variables present in the cluster. On the other hand, the mean for the observations of the dataset for the variable Education is like the mean for the observations of the cluster.

This indicates that the clusters Number of Companies Worked, and Job Level are some of the lowest contributing factors to HR attrition in a company.

Data Modeling

Data Partition

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Filter_TRAIN	1387
TRAIN	EMWS1.Part_TRAIN	832
VALIDATE	EMWS1.Part_VALIDATE	278
TEST	EMWS1.Part_TEST	277

The data was partitioned into 60:20:20. 60 for training data, 20 for validation data and 20 for test data. Total observations 1387. Training set has 832 and the Validation set has 278 and the Test set has 277 after partition. We chose a higher percentage for the training data to get more accurate predictions as needed in the analysis.

Logistic Regression

Identification of Key Metrics:
Key Measures used

Goodness of Fit

p value for testing the significance of a predictor entered in the model.

To test the impact of IV'S on DV

Odds Ratio

Project Scope

- Performing cluster analysis (as a baseline model), decision trees, random forest, logistic regression (stepwise, backward, and forward) to determine the best model for predicting attrition.
- Find trends by modeling relationships between attrition and variables in bar charts.
- Providing summary statistics about employees.
- Understanding how the company's employees perceive the working environment.
- Investigating possible factors that affect attrition.
- Providing insight to managers.

Model 1 Stepwise Regression

Likelihood Ratio Test:

The Likelihood Ratio Test for the model was significant or indicated a good fit for the model.

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
744.345	472.016	272.3290	25	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	4.4483	0.0349
BusinessTravel	2	12.0094	0.0025
DistanceFromHome	1	11.2502	0.0008
EnvironmentSatisfaction	1	15.9098	<.0001
Gender	1	4.3358	0.0373
JobInvolvement	1	14.1445	0.0002
JobRole	8	21.0320	0.0071
JobSatisfaction	1	14.7596	0.0001
MaritalStatus	2	21.6918	<.0001
NumCompaniesWorked	1	18.5427	<.0001
OverTime	1	58.4586	<.0001
RelationshipSatisfaction	1	13.3038	0.0003
TotalWorkingYears	1	6.7351	0.0095
TrainingTimesLastYear	1	6.9337	0.0085
YearsInCurrentRole	1	8.5684	0.0034
YearsSinceLastPromotion	1	14.8765	0.0001

The analysis of the effects table indicates that the variables age, business travel, distance from home, environment satisfaction, gender, job involvement, job role, job satisfaction, marital status, number of companies worked, overtime, relationship satisfaction, total working years, training times last year, years in current role and years since last promotion are significant predictors of employee attrition.

Output

Analysis of Maximum Likelihood Estimates

Parameter		Attrition	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		Yes	1	2.8141	22.2721	0.02	0.8995		16.679
Age		Yes	1	-0.0388	0.0184	4.45	0.0349	-0.1887	0.962
BusinessTravel	Non-Travel	Yes	1	-0.7377	0.3052	5.84	0.0156		0.478
BusinessTravel	Travel_Frequently	Yes	1	0.8040	0.2338	11.83	0.0006		2.234
DistanceFromHome		Yes	1	0.0487	0.0145	11.25	0.0008	0.2160	1.050
EnvironmentSatisfaction		Yes	1	-0.4317	0.1082	15.91	<.0001	-0.2617	0.649
Gender	Female	Yes	1	-0.2685	0.1289	4.34	0.0373		0.765
JobInvolvement		Yes	1	-0.5958	0.1584	14.14	0.0002	-0.2364	0.551
JobRole	Healthcare Representative	Yes	1	1.4308	22.2575	0.00	0.9487		4.182
JobRole	Human Resources	Yes	1	3.3532	22.2566	0.02	0.8802		28.595
JobRole	Laboratory Technician	Yes	1	2.6363	22.2540	0.01	0.9057		13.962
JobRole	Manager	Yes	1	-7.9819	116.7	0.00	0.9455		0.000
JobRole	Manufacturing Director	Yes	1	1.5452	22.2567	0.00	0.9447		4.689
JobRole	Research Director	Yes	1	-8.6711	136.3	0.00	0.9493		0.000
JobRole	Research Scientist	Yes	1	2.0162	22.2541	0.01	0.9278		7.510
JobRole	Sales Executive	Yes	1	2.2205	22.2539	0.01	0.9205		9.212
JobSatisfaction		Yes	1	-0.4219	0.1098	14.76	0.0001	-0.2548	0.656
MaritalStatus	Divorced	Yes	1	-0.4540	0.2091	4.72	0.0299		0.635
MaritalStatus	Married	Yes	1	-0.3339	0.1703	3.84	0.0500		0.716
NumCompaniesWorked		Yes	1	0.2151	0.0499	18.54	<.0001	0.2992	1.240
OverTime	No	Yes	1	-1.0059	0.1316	58.46	<.0001		0.366
RelationshipSatisfaction		Yes	1	-0.3984	0.1092	13.30	0.0003	-0.2396	0.671
TotalWorkingYears		Yes	1	-0.0895	0.0345	6.74	0.0095	-0.3321	0.914
TrainingTimesLastYear		Yes	1	-0.2681	0.1018	6.93	0.0085	-0.1910	0.765
YearsInCurrentRole		Yes	1	-0.1625	0.0555	8.57	0.0034	-0.2908	0.850
YearsSinceLastPromotion		Yes	1	0.2283	0.0592	14.88	0.0001	0.3149	1.257

The analysis of the maximum likelihood estimates table is an indicator of how different levels of a categorical variable could be non-significant even if those variables are significant overall. For example, Job role which was significant overall is non-significant at certain levels like healthcare representative, human resources laboratory technician, etc. Similarly, Marital Status which was significant shows significance for only divorced people and the married people show an insignificant impact on the analysis of maximum likelihood estimates table.

Results - Node: Stepwise Logistic Regression (2) Diagram: IBM

File Edit View Window

Output Plot...

Odds Ratio Estimates

Effect		Attrition	Point Estimate
Age		Yes	0.962
BusinessTravel	Non-Travel vs Travel_Rarely	Yes	0.511
BusinessTravel	Travel_Frequently vs Travel_Rarely	Yes	2.388
DistanceFromHome		Yes	1.050
EnvironmentSatisfaction		Yes	0.649
Gender	Female vs Male	Yes	0.585
JobInvolvement		Yes	0.551
JobRole	Healthcare Representative vs Sales Representative	Yes	0.133
JobRole	Human Resources vs Sales Representative	Yes	0.907
JobRole	Laboratory Technician vs Sales Representative	Yes	0.443
JobRole	Manager vs Sales Representative	Yes	<0.001
JobRole	Manufacturing Director vs Sales Representative	Yes	0.149
JobRole	Research Director vs Sales Representative	Yes	<0.001
JobRole	Research Scientist vs Sales Representative	Yes	0.238
JobRole	Sales Executive vs Sales Representative	Yes	0.292
JobSatisfaction		Yes	0.656
MaritalStatus	Divorced vs Single	Yes	0.289
MaritalStatus	Married vs Single	Yes	0.326
NumCompaniesWorked		Yes	1.240
OverTime	No vs Yes	Yes	0.134
RelationshipSatisfaction		Yes	0.671
TotalWorkingYears		Yes	0.914
TrainingTimesLastYear		Yes	0.765
YearsInCurrentRole		Yes	0.850
YearsSinceLastPromotion		Yes	1.257

Interpreting the odds ratio for Model 1:

- As the age of the employee increases, the odds of the attrition of an employee decreases by 0.962.
- The employees who do not travel frequently have 0.511 lower odds of attrition than employees who travel rarely.
- The employees who travel frequently have higher odds of attrition than employees who travel rarely. The odds of attrition for such employees are 2.388 times that of employees who travel rarely.
- A unit increase in the distance from home to workplace for an employee increases the odds of attrition of an employee 1.050 times.
- A unit increase in environment satisfaction for an employee decreases the odds of attrition of an employee by 0.649 times.
- The female employees have 0.585 lower odds of attrition than male employees.
- A unit increase in job involvement of an employee decreases the odds of attrition of an employee 0.551 times.

- A unit increase in job satisfaction for an employee decreases the odds of attrition of an employee by 0.656 times.
- The employees who have "divorced" as their marital status have 0.289 lower odds of attrition than employees who have "single" as their marital status.
- A unit increase in the number of companies an employee works with increases the odds of attrition of an employee 1.240 times.
- The employees who do not work overtime have 0.134 times lower odds of attrition than those employees who work overtime.
- A unit increase in the relationship satisfaction of an employee lowers the odds of attrition of an employee 0.671 times.
- A unit increase in the total working years of an employee lowers the odds of attrition of an employee 0.914 times.
- A unit increase in the number of training times last year for an employee lowers the odds of attrition of an employee 0.765 times.
- A unit increase in the number of years an employee spent in a current role lowers the odds of attrition of an employee 0.850 times.
- A unit increase in the number of years an employee spends after last promotion increases the odds of attrition of an employee 1.257 times.

MODEL 2 (Forward Logistic Regression)

Likelihood Ratio Test

The Likelihood Ratio Test for the model was significant or indicated a good fit for the model.

Results - Node: Backward Logistic Regression (3) Diagram: IBM

File Edit View Window

Output

Convergence criterion (GCONV=1E-6) satisfied.

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
744.345	465.778	278.5667	27	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	4.4483	0.0349
BusinessTravel	2	12.0094	0.0025
DistanceFromHome	1	11.2502	0.0008
EnvironmentSatisfaction	1	15.9098	<.0001
Gender	1	4.3358	0.0373
JobInvolvement	1	14.1445	0.0002
JobRole	8	21.0320	0.0071
JobSatisfaction	1	14.7596	0.0001
MaritalStatus	2	21.6918	<.0001
NumCompaniesWorked	1	18.5427	<.0001
OverTime	1	58.4586	<.0001
RelationshipSatisfaction	1	13.3038	0.0003
TotalWorkingYears	1	6.7351	0.0095
TrainingTimesLastYear	1	6.9337	0.0085
YearsInCurrentRole	1	8.5684	0.0034
YearsSinceLastPromotion	1	14.8765	0.0001

The analysis of the effects table indicates that the variables age, business travel, distance from home, environment satisfaction, gender, job involvement, job role, job satisfaction, marital status, number of companies worked, overtime, relationship satisfaction, total working years, training times last year, years in current role and years since last promotion are significant predictors of employee attrition.

Results - Node: Forward Logistic Regression Diagram: IBM

File Edit View Window

Output

Analysis of Maximum Likelihood Estimates

Parameter		Attrition	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		Yes	1	2.8141	22.2721	0.02	0.8995		16.679
Age		Yes	1	-0.0388	0.0184	4.45	0.0349	-0.1887	0.962
BusinessTravel	Non-Travel	Yes	1	-0.7377	0.3052	5.84	0.0156		0.478
BusinessTravel	Travel_Frequently	Yes	1	0.8040	0.2338	11.83	0.0006		2.234
DistanceFromHome		Yes	1	0.0487	0.0145	11.25	0.0008	0.2160	1.050
EnvironmentSatisfaction		Yes	1	-0.4317	0.1082	15.91	<.0001	-0.2617	0.649
Gender	Female	Yes	1	-0.2685	0.1289	4.34	0.0373		0.765
JobInvolvement		Yes	1	-0.5958	0.1584	14.14	0.0002	-0.2364	0.551
JobRole	Healthcare Representative	Yes	1	1.4308	22.2575	0.00	0.9487		4.182
JobRole	Human Resources	Yes	1	3.3532	22.2566	0.02	0.8802		28.595
JobRole	Laboratory Technician	Yes	1	2.6363	22.2540	0.01	0.9057		13.962
JobRole	Manager	Yes	1	-7.9819	116.7	0.00	0.9455		0.000
JobRole	Manufacturing Director	Yes	1	1.5452	22.2567	0.00	0.9447		4.689
JobRole	Research Director	Yes	1	-8.6711	136.3	0.00	0.9493		0.000
JobRole	Research Scientist	Yes	1	2.0162	22.2541	0.01	0.9278		7.510
JobRole	Sales Executive	Yes	1	2.2205	22.2539	0.01	0.9205		9.212
JobSatisfaction		Yes	1	-0.4219	0.1098	14.76	0.0001	-0.2548	0.656
MaritalStatus	Divorced	Yes	1	-0.4540	0.2091	4.72	0.0299		0.635
MaritalStatus	Married	Yes	1	-0.3339	0.1703	3.84	0.0500		0.716
NumCompaniesWorked		Yes	1	0.2151	0.0499	18.54	<.0001	0.2992	1.240
OverTime	No	Yes	1	-1.0059	0.1316	58.46	<.0001		0.366
RelationshipSatisfaction		Yes	1	-0.3984	0.1092	13.30	0.0003	-0.2396	0.671
TotalWorkingYears		Yes	1	-0.0895	0.0345	6.74	0.0095	-0.3321	0.914
TrainingTimesLastYear		Yes	1	-0.2681	0.1018	6.93	0.0085	-0.1910	0.765
YearsInCurrentRole		Yes	1	-0.1625	0.0555	8.57	0.0034	-0.2908	0.850
YearsSinceLastPromotion		Yes	1	0.2283	0.0592	14.88	0.0001	0.3149	1.257

The analysis of the maximum likelihood estimates table is an indicator of how different levels of a categorical variable could be non-significant even if those variables are significant overall. For example, Job role which was significant overall is non-significant at certain levels like healthcare representative, human resources laboratory technician etc. Similarly, marital status which was significant shows significance for only divorced people and the married people show an insignificant impact on the analysis of maximum likelihood estimates table.

Results - Node: Forward Logistic Regression Diagram: IBM

File Edit View Window

Output

Odds Ratio Estimates

Effect		Attrition	Point Estimate
Age		Yes	0.962
BusinessTravel	Non-Travel vs Travel_Rarely	Yes	0.511
BusinessTravel	Travel_Frequently vs Travel_Rarely	Yes	2.388
DistanceFromHome		Yes	1.050
EnvironmentSatisfaction		Yes	0.649
Gender	Female vs Male	Yes	0.585
JobInvolvement		Yes	0.551
JobRole	Healthcare Representative vs Sales Representative	Yes	0.133
JobRole	Human Resources vs Sales Representative	Yes	0.907
JobRole	Laboratory Technician vs Sales Representative	Yes	0.443
JobRole	Manager vs Sales Representative	Yes	<0.001
JobRole	Manufacturing Director vs Sales Representative	Yes	0.149
JobRole	Research Director vs Sales Representative	Yes	<0.001
JobRole	Research Scientist vs Sales Representative	Yes	0.238
JobRole	Sales Executive vs Sales Representative	Yes	0.292
JobSatisfaction		Yes	0.656
MaritalStatus	Divorced vs Single	Yes	0.289
MaritalStatus	Married vs Single	Yes	0.326
NumCompaniesWorked		Yes	1.240
OverTime	No vs Yes	Yes	0.134
RelationshipSatisfaction		Yes	0.671
TotalWorkingYears		Yes	0.914
TrainingTimesLastYear		Yes	0.765
YearsInCurrentRole		Yes	0.850
YearsSinceLastPromotion		Yes	1.257

Interpreting the odds ratio for Model 2:

- As the age of the employee increases, the odds of the attrition of an employee decreases by 0.962
- The employees who do not travel frequently have 0.511 lower odds of attrition than employees who travel rarely.
- The employees who travel frequently have higher odds of attrition than employees who travel rarely. The odds of attrition for such employees are 2.388 times that of employees who travel rarely.
- A unit increase in the distance from home to workplace for an employee increases the odds of attrition of an employee 1.050 times.
- A unit increase in environment satisfaction for an employee decreases the odds of attrition of an employee by 0.649 times.
- The female employees have 0.585 lower odds of attrition than male employees.
- A unit increase in job involvement of an employee decreases the odds of attrition of an employee 0.551 times.
- A unit increase in job satisfaction for an employee decreases the odds of attrition of an employee by 0.656 times.
- The employees who have “divorced” as their marital status have 0.289 lower odds of attrition than employees who have “single” as their marital status.

- A unit increase in the number of companies an employee works with increases the odds of attrition of an employee 1.240 times.
- The employees who do not work overtime have 0.134 times lower odds of attrition than those employees who work overtime.
- A unit increase in the relationship satisfaction of an employee lowers the odds of attrition of an employee 0.671 times.
- A unit increase in the total working years of an employee lowers the odds of attrition of an employee 0.914 times.
- A unit increase in the number of training times last year for an employee lowers the odds of attrition of an employee 0.765 times.
- A unit increase in the number of years an employee spent in a current role lowers the odds of attrition of an employee 0.850 times.
- A unit increase in the number of years an employee spends after last promotion increases the odds of attrition of an employee 1.257 times.

MODEL 3 (Backward logistic regression)

Likelihood Ratio Test

The Likelihood Ratio Test for the model was significant or indicated a good fit for the model.

Results - Node: Backward Logistic Regression (3) Diagram: IBM

File Edit View Window

Output

```
Convergence criterion (GCONV=1E-6) satisfied.


     Likelihood Ratio Test for Global Null Hypothesis: BETA=0

     -2 Log Likelihood          Likelihood
  Intercept     Intercept &          Ratio
       Only      Covariates     Chi-Square       DF     Pr > ChiSq

    744.345         465.778       278.5667       27         <.0001
```

Results - Node: Backward Logistic Regression (3) Diagram: IBM

File Edit View Window

Output

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	4.4483	0.0349
BusinessTravel	2	12.0094	0.0025
DistanceFromHome	1	11.2502	0.0008
EnvironmentSatisfaction	1	15.9098	<.0001
Gender	1	4.3358	0.0373
JobInvolvement	1	14.1445	0.0002
JobRole	8	21.0320	0.0071
JobSatisfaction	1	14.7596	0.0001
MaritalStatus	2	21.6918	<.0001
NumCompaniesWorked	1	18.5427	<.0001
OverTime	1	58.4586	<.0001
RelationshipSatisfaction	1	13.3038	0.0003
TotalWorkingYears	1	6.7351	0.0095
TrainingTimesLastYear	1	6.9337	0.0085
YearsInCurrentRole	1	8.5684	0.0034
YearsSinceLastPromotion	1	14.8765	0.0001

The analysis of the effects table indicates that the variables age, business travel, distance from home, environment satisfaction, gender, job involvement, job role, job satisfaction, marital status, number of companies worked, overtime, relationship satisfaction, total working years, training times last year, years in current role and years since last promotion are significant predictors of employee attrition.

Results - Node: Backward Logistic Regression (3) Diagram: IBM

File Edit View Window

Output

Analysis of Maximum Likelihood Estimates

Parameter		Attrition	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		Yes	1	2.8141	22.2721	0.02	0.8995		16.679
Age		Yes	1	-0.0388	0.0184	4.45	0.0349	-0.1887	0.962
BusinessTravel	Non-Travel	Yes	1	-0.7377	0.3052	5.84	0.0156		0.478
BusinessTravel	Travel_Frequently	Yes	1	0.8040	0.2338	11.83	0.0006		2.234
DistanceFromHome		Yes	1	0.0487	0.0145	11.25	0.0008	0.2160	1.050
EnvironmentSatisfaction		Yes	1	-0.4317	0.1082	15.91	<.0001	-0.2617	0.649
Gender	Female	Yes	1	-0.2685	0.1289	4.34	0.0373		0.765
JobInvolvement		Yes	1	-0.5958	0.1584	14.14	0.0002	-0.2364	0.551
JobRole	Healthcare Representative	Yes	1	1.4308	22.2575	0.00	0.9487		4.182
JobRole	Human Resources	Yes	1	3.3532	22.2566	0.02	0.8802		28.595
JobRole	Laboratory Technician	Yes	1	2.6363	22.2540	0.01	0.9057		13.962
JobRole	Manager	Yes	1	-7.9819	116.7	0.00	0.9455		0.000
JobRole	Manufacturing Director	Yes	1	1.5452	22.2567	0.00	0.9447		4.689
JobRole	Research Director	Yes	1	-8.6711	136.3	0.00	0.9493		0.000
JobRole	Research Scientist	Yes	1	2.0162	22.2541	0.01	0.9278		7.510
JobRole	Sales Executive	Yes	1	2.2205	22.2539	0.01	0.9205		9.212
JobSatisfaction		Yes	1	-0.4219	0.1098	14.76	0.0001	-0.2548	0.656
MaritalStatus	Divorced	Yes	1	-0.4540	0.2091	4.72	0.0299		0.635
MaritalStatus	Married	Yes	1	-0.3339	0.1703	3.84	0.0500		0.716
NumCompaniesWorked		Yes	1	0.2151	0.0499	18.54	<.0001	0.2992	1.240
OverTime	No	Yes	1	-1.0059	0.1316	58.46	<.0001		0.366
RelationshipSatisfaction		Yes	1	-0.3984	0.1092	13.30	0.0003	-0.2396	0.671
TotalWorkingYears		Yes	1	-0.0895	0.0345	6.74	0.0095	-0.3321	0.914
TrainingTimesLastYear		Yes	1	-0.2681	0.1018	6.93	0.0085	-0.1910	0.765
YearsInCurrentRole		Yes	1	-0.1625	0.0555	8.57	0.0034	-0.2908	0.850
YearsSinceLastPromotion		Yes	1	0.2283	0.0592	14.88	0.0001	0.3149	1.257

The analysis of the maximum likelihood estimates table is an indicator of how different levels of a categorical variable could be non-significant, even if those variables are significant overall. For example, Job role which was significant overall is non-significant at certain levels like healthcare representative, human resources laboratory technician etc. Similarly, marital status which was significant, shows significance for only divorced people and the married people show an insignificant impact on the analysis of maximum likelihood estimates table.

Interpreting the odds ratio (Model 3)

- As the age of the employee increases, the odds of the attrition of an employee decreases by 0.962
- The employees who do not travel frequently have 0.511 lower odds of attrition than employees who travel rarely.
- The employees who travel frequently have higher odds of attrition than employees who travel rarely. The odds of attrition for such employees are 2.388 times that of employees who travel rarely.
- A unit increase in the distance from home to workplace for an employee increases the odds of attrition of an employee 1.050 times.
- A unit increase in environment satisfaction for an employee decreases the odds of attrition of an employee by 0.649 times.
- The female employees have 0.585 lower odds of attrition than male employees.
- A unit increase in job involvement of an employee decreases the odds of attrition of an employee 0.551 times.
- A unit increase in job satisfaction for an employee decreases the odds of attrition of an employee by 0.656 times.
- The employees who have divorced as their marital status have 0.289 lower odds of attrition than employees who have single as their marital status.
- A unit increase in the number of companies an employee works with increases the odds of attrition of an employee 1.240 times.
- The employees who do not work overtime have 0.134 times lower odds of attrition than those employees who work overtime.

- A unit increase in the relationship satisfaction of an employee lowers the odds of attrition of an employee 0.671 times.
- A unit increase in the total working years of an employee lowers the odds of attrition of an employee 0.914 times.
- A unit increase in the number of training times last year for an employee lowers the odds of attrition of an employee 0.765 times.
- A unit increase in the number of years an employee spent in a current role lowers the odds of attrition of an employee 0.850 times.
- A unit increase in the number of years an employee spends after last promotion increases the odds of attrition of an employee 1.257 times.

Model comparison (Logistic Regression)

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg	Forward Logistic Regression	0.12590	0.082139	0.10337	0.098461
	Reg2	Stepwise Logistic Regression	0.12590	0.082139	0.10337	0.098461
	Reg3	Backward Logistic Regression	0.12590	0.082139	0.10337	0.098461

All three regression methods (stepwise, backward, and forward) lead to similar variables which showed a significant impact for the variable attrition.

The impact was similar for all the models in terms of direction as well as magnitude.

Random Forest

Implementing Machine Learning models:

We implemented a Random Forest after which we looked at feature importance from these respective models. Random Forests are a data mining algorithm that can select important variables. In a random forest, the target variable can be categorical or quantitative. The forest is used to rank the importance of variables in predicting a target.

Analysis and Interpretation of Results:

The dataset is imbalanced, the Random Forest classifier in SAS Miner also contains a very convenient attribute feature importance which tells us which features within our dataset have been given most importance using Random Forest. Shown below is a diagram of the various feature importance.

Number of Observations

Type	NTrain	NValid	NTotal
Number of Observations Read	832	278	1110
Number of Observations Used	832	278	1110

Baseline Fit Statistics

Statistic	Value	Validation
Average Square Error	0.138	0.138
Misclassification Rate	0.165	0.165
Log Loss	0.447	0.449

```
Model Events

                                          Number
                          Measurement       of
 Target       Event          Level        Levels       Order        Label

Attrition      YES         NOMINAL          2       Descending



Predicted and decision variables

Type          Variable            Label

TARGET        Attrition
PREDICTED     P_AttritionYes      Predicted: Attrition=Yes
RESIDUAL      R_AttritionYes      Residual: Attrition=Yes
PREDICTED     P_AttritionNo       Predicted: Attrition=No
RESIDUAL      R_AttritionNo       Residual: Attrition=No
FROM          F_Attrition         From: Attrition
INTO          I_Attrition         Into: Attrition



The HPFOREST Procedure

      Performance Information

Execution Mode       Single-Machine
Number of Threads    2
```

The data read and used is 832, the misclassification rate is 0.165% to interpret it correctly, classified approximately 0.84 % of the sample.

Further analyzing the fit statistics from the above data, hp forest computes fit statistics in the 1st 16 observations for a sequence of forest that have an increasing number of trees. As the number of trees increases, the fit statistics improve, but here it remains stable after the 12th observation and it remains stable for long in the data. There are not many fluctuations that can be observed in a small range. The forest model provides an alternate estimate of Average square error and misclassification rate (Train and Out of Bag). The OOB estimate is the convenient estimate that is based on test data and has a less biased estimate about how the model will perform.

Fit Statistics

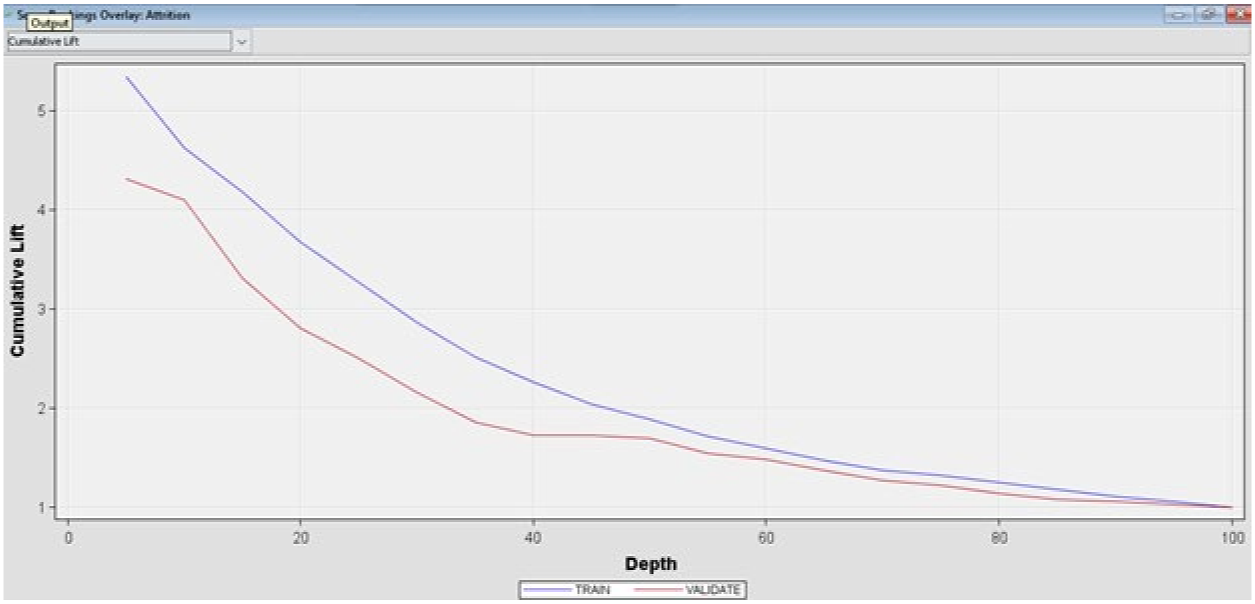
Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (OOB)	Average Square Error (Valid)	Misclassification Rate (Train)	Misclassification Rate (OOB)	Misclassification Rate (Valid)	Log Loss (Train)	Log Loss (OOB)	Log Loss (Valid)
1	7	0.124	0.151	0.139	0.167	0.207	0.183	0.439	0.591	0.536
2	19	0.111	0.139	0.135	0.141	0.175	0.169	0.361	0.566	0.455
3	29	0.106	0.135	0.131	0.148	0.172	0.162	0.346	0.536	0.423
4	34	0.106	0.128	0.126	0.157	0.162	0.158	0.348	0.450	0.408
5	44	0.106	0.123	0.123	0.157	0.162	0.158	0.347	0.430	0.401
6	61	0.102	0.117	0.121	0.147	0.155	0.158	0.338	0.412	0.397
7	72	0.101	0.117	0.120	0.143	0.153	0.158	0.336	0.409	0.395
8	87	0.100	0.118	0.120	0.144	0.155	0.155	0.333	0.387	0.393
9	94	0.102	0.117	0.120	0.154	0.159	0.155	0.337	0.384	0.395
10	106	0.101	0.118	0.120	0.160	0.156	0.155	0.335	0.384	0.398
11	112	0.102	0.118	0.120	0.157	0.159	0.155	0.337	0.384	0.396
12	118	0.103	0.119	0.119	0.159	0.160	0.158	0.339	0.387	0.395
13	124	0.103	0.119	0.119	0.155	0.160	0.158	0.342	0.388	0.394
14	139	0.102	0.120	0.118	0.154	0.159	0.155	0.337	0.389	0.393
15	143	0.102	0.119	0.118	0.148	0.160	0.151	0.338	0.387	0.393
16	149	0.103	0.119	0.118	0.150	0.160	0.151	0.340	0.388	0.391

Variable Importance

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction	Label
OverTime	82	0.010120	0.020239	0.00686	0.01703	0.00329	0.01363	
JobLevel	80	0.007678	0.015356	0.00496	0.01258	0.00412	0.01628	
StockOptio...	67	0.004845	0.009689	0.00193	0.00655	0.00435	0.01092	
TotalWorkin...	37	0.004061	0.008121	0.00072	0.00436	0.00311	0.00880	
MaritalStatus	36	0.002497	0.004993	0.00053	0.00332	0.00161	0.00441	
Department	34	0.001287	0.002574	-0.00050	0.00077	-0.00059	0.00018	
JobRole	34	0.002998	0.005995	-0.00059	0.00268	0.00048	0.00359	
Relationshi...	34	0.001275	0.002549	-0.00019	0.00136	-0.00200	0.00004	
Environme...	30	0.001673	0.003347	-0.00060	0.00103	0.00044	0.00138	
EducationFi...	28	0.001237	0.002473	-0.00015	0.00158	-0.00087	0.00016	
JobInvolve...	28	0.001551	0.003101	-0.00021	0.00140	-0.00076	0.00123	
WorkLifeBa...	25	0.000983	0.001966	-0.00064	0.00015	-0.00071	0.00018	
YearsAtCo...	25	0.002139	0.004278	-0.00047	0.00136	0.00047	0.00232	
MonthlyInco...	19	0.002241	0.004482	-0.00068	0.00160	0.00046	0.00308	
YearsWithC...	18	0.001096	0.002191	-0.00059	0.00052	0.00063	0.00143	
YearsSince...	17	0.000550	0.001099	-0.00067	-0.00007	-0.00026	0.00018	
TrainingTi...	16	0.000870	0.001741	-0.00062	0.00032	-0.00091	-0.00004	
VAR1	16	0.001488	0.002976	-0.00021	0.00135	0.00017	0.00255	ï»¿Age
YearsInCur...	16	0.001179	0.002358	-0.00015	0.00114	0.00068	0.00185	
BusinessTr...	15	0.000810	0.001620	-0.00059	0.00036	-0.00062	0.00010	
EmployeeN...	15	0.000595	0.001190	-0.00044	-0.00008	-0.00100	-0.00024	
JobSatisfac...	14	0.000785	0.001570	-0.00014	0.00073	0.00001	0.00076	
DistanceFr...	13	0.001059	0.002117	-0.00072	0.00058	-0.00046	0.00045	
Performanc...	13	0.000469	0.000939	-0.00018	0.00023	-0.00046	0.00007	
NumComp...	11	0.000791	0.001581	-0.00039	0.00045	-0.00139	-0.00064	
Education	7	0.000330	0.000660	-0.00031	0.00011	-0.00074	-0.00055	
MonthlyRate	6	0.000183	0.000365	-0.00013	0.00009	-0.00022	-0.00010	
PercentSal...	6	0.000381	0.000762	-0.00033	0.00015	-0.00047	-0.00013	
DailyRate	5	0.000347	0.000695	-0.00038	-0.00009	-0.00028	0.00008	
Gender	4	0.000206	0.000413	-0.00041	-0.00011	-0.00015	0.00004	
HourlyRate	3	0.000192	0.000383	0.00008	0.00026	-0.00001	0.00020	
StandardH...	0	0.000000	0.000000	0.00000	0.00000	0.00000	0.00000	

The arguably largest contribution in the random forest output is from Loss Reduction Variable Importance, specifically the variable importance ranking. As for the fit statistics, the OOB (Out of Bag) data is less biased. The variables are listed from higher importance to lower in predicting data.

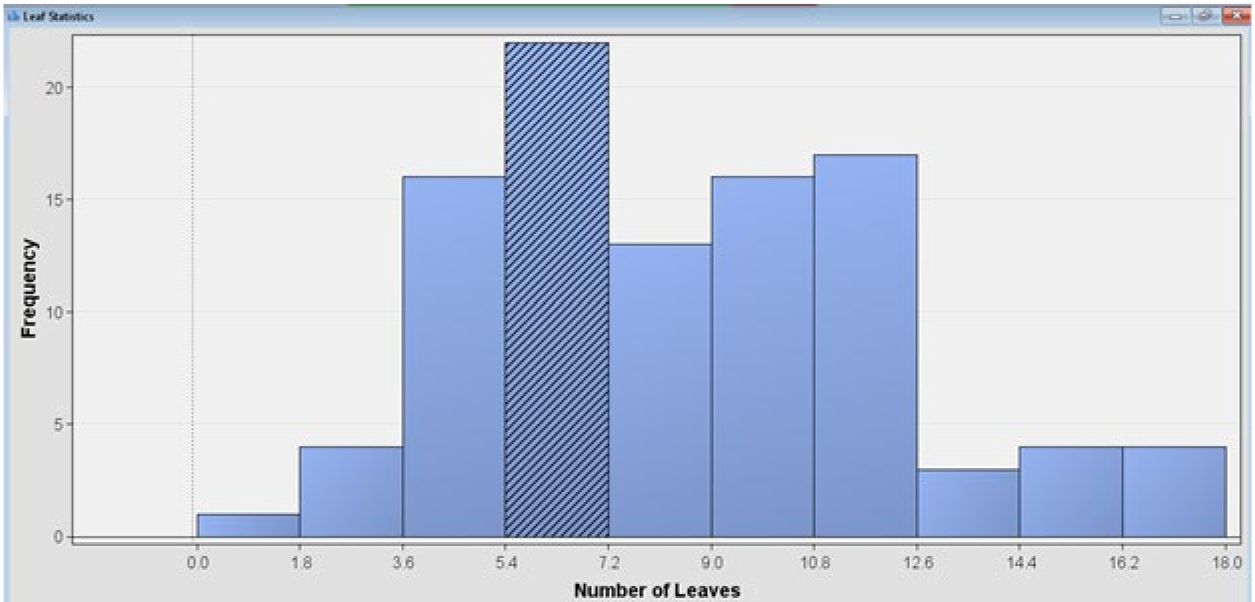
It plots a significant difference between train - out of bag validation curves. The model has a bumpy graph all through. The out of bag misclassification is used as the assessment statistics because it represents the accuracy that is expected in a general independent dataset. There is not much accuracy gained in Training and Out of bag graphs, it is not a level off position at the end for the Out of bag graph. The data has a stable and bumpy graph for validation and out of bag data too shows a picture like that. More bumpy graph observations can be seen for Training data.

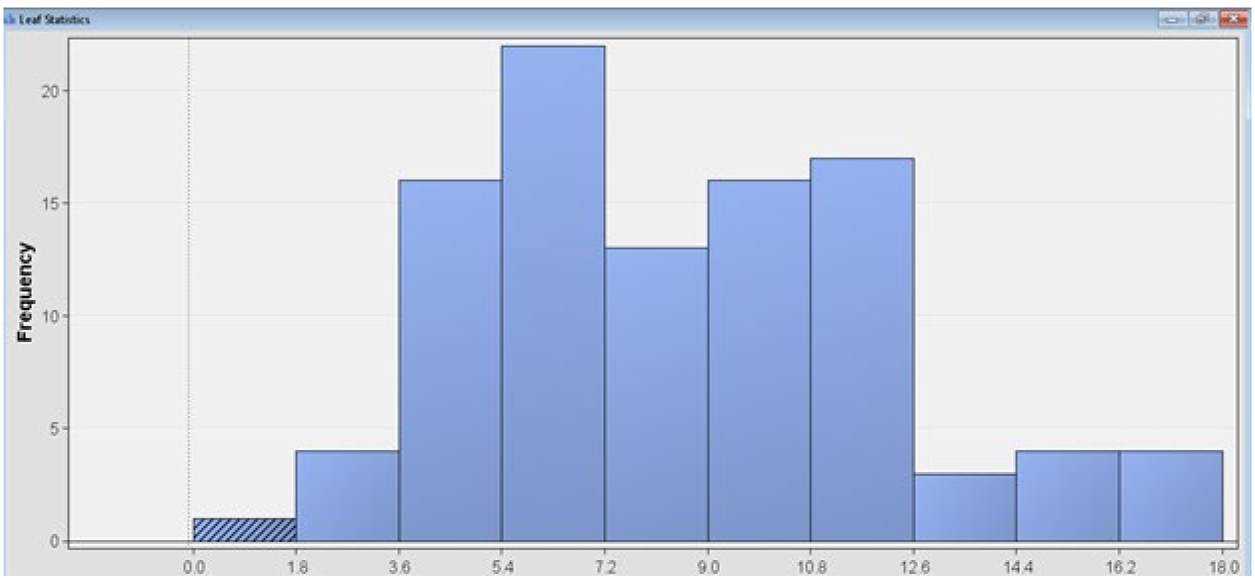


The validation and training data shows a bumpy picture until depth 40 where the cumulative lift is 1.726 for validation, similarly, for the training data set, the cumulative lift is 2.26.

Tree Assessment:

Leaf Statistics analysis for 13 leaves are as follows: the highest number of leaves is between 5.4 to 7.2, where the frequency observed is 22 which has been highlighted below. Similarly lowest observations have been recorded between 0 to 1.8 (frequency - 1).





Iteration History:

The analysis for a small sample of data has been done, starting from 9 trees. It states that the misclassification rate (Validate) remains between 0.151 to 0.158. Similarly, the misclassification rate (Out of Bag) remains between 0.155 to 0.160. Also, the misclassification rate (Train) remains between 0.154 to 0.160. Much variation has been observed in the data.

Iteration History

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)	Misclassification Rate (Train)	Misclassification Rate (Out of Bag)	Misclassification Rate (Validate)	Log Loss (Train)	Log Loss (Out of Bag)	Log Loss (Validate)
1	7	0.124	0.151	0.139	0.167	0.207	0.183	0.439	0.591	0.536
2	19	0.111	0.139	0.135	0.141	0.175	0.169	0.361	0.566	0.455
3	29	0.106	0.135	0.131	0.148	0.172	0.162	0.346	0.536	0.423
4	34	0.106	0.128	0.126	0.157	0.162	0.158	0.348	0.450	0.408
5	44	0.106	0.123	0.123	0.157	0.162	0.158	0.347	0.430	0.401
6	61	0.102	0.117	0.121	0.147	0.155	0.158	0.338	0.412	0.397
7	72	0.101	0.117	0.120	0.143	0.153	0.158	0.336	0.409	0.395
8	87	0.100	0.118	0.120	0.144	0.155	0.155	0.333	0.387	0.393
9	94	0.102	0.117	0.120	0.154	0.159	0.155	0.337	0.384	0.395
10	106	0.101	0.118	0.120	0.160	0.156	0.155	0.335	0.384	0.398
11	112	0.102	0.118	0.120	0.157	0.159	0.155	0.337	0.384	0.396
12	118	0.103	0.119	0.119	0.159	0.160	0.158	0.339	0.387	0.395
13	124	0.103	0.119	0.119	0.155	0.160	0.158	0.342	0.388	0.394
14	139	0.102	0.120	0.118	0.154	0.159	0.155	0.337	0.389	0.393
15	143	0.102	0.119	0.118	0.148	0.160	0.151	0.338	0.387	0.393
16	149	0.103	0.119	0.118	0.150	0.160	0.151	0.340	0.388	0.391

Based on the above results, a random forest model gave us a reasonable result, in terms of accuracy and specificity.

The model gave an accuracy score of 0.84, which is not too bad. The random forest works quite well even with the default parameters. This can be improved though by tuning hyperparameters of the Random Forest classifier. Random forest also does not over fit easily because of its randomness feature.

Decision Trees

Since, the target variable is a category, we decided to apply the Decision Tree to find out the attrition rate.

Decision Trees are a nonparametric supervised learning method used for classification and regression. It is also referred to as CART: Classification and Regression Tree. In simpler words, it is a graphical representation of all the possible solutions to a decision based on certain conditions. Tree models where the target variable can take a finite set of values are called classification trees and target variables can take continuous values (numbers) are called regression trees. (Ref: towardsdatascience.com)

We built two models autonomously and one model interactively. Then, we evaluated all the three models using model comparison.

For Decision tree – Model 1, we used the default settings and built the model.

Property	Value
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

For Decision tree – Model 2, we changed the maximum branch into 3 and built the model.

Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	3
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

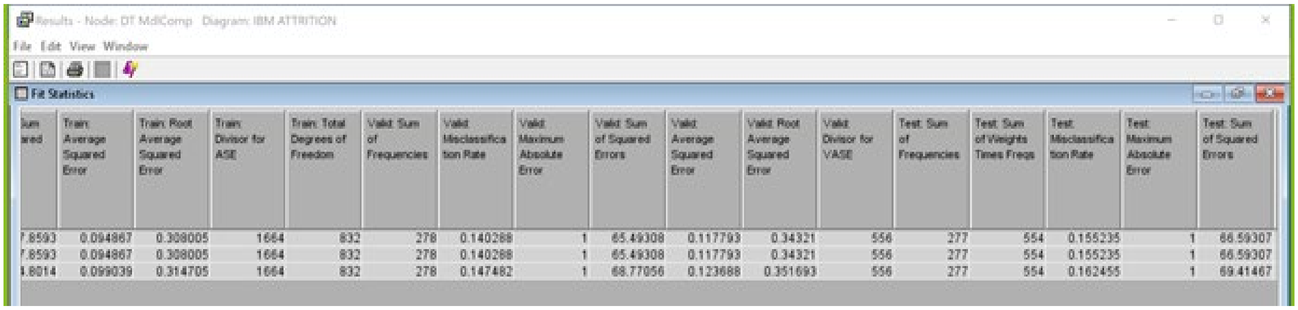
For Decision Tree – Model 3 we changed the maximum branch into 5 and built the model.

Property	Value
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	5
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No

For all the above models, we have selected an assessment measure to "Misclassification rate" in the respective model properties in SAS EM as our target variable.

Model Comparison (Fit Statistics)

To assess the accuracy of each model, we evaluated each model against the validation dataset using the "Average Square Error" and "Root Mean Square Error" and misclassification rate metric. We created all the three models and connected it to the model comparison node in SAS Miner. Based on the result of the Decision Tree, Model 1 came out as the best model compared to all other models. Although Model 2 has similar values with Model 1, Model 1 was chosen based on its simplicity (2 branches compared to 3 branches in Model 2).



Results - Node: DT MdlComp Diagram: IBM ATTRITION

File Edit View Window

Fit Statistics

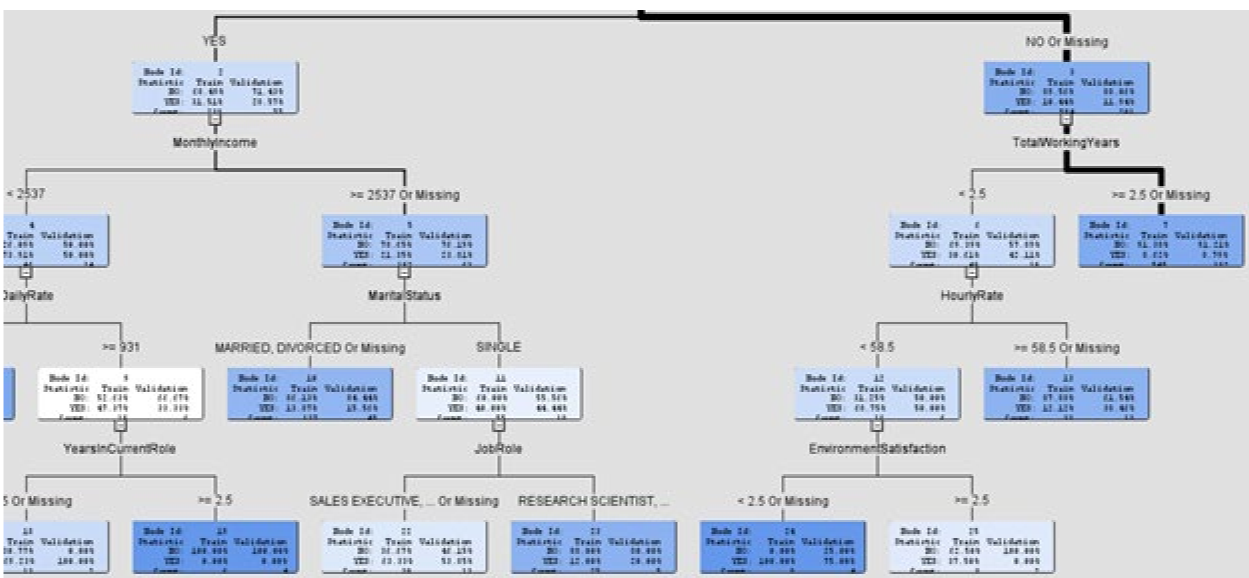
...um ...red	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Test: Sum of Frequencies	Test: Sum of Weights Times Freqs	Test: Misclassification Rate	Test: Maximum Absolute Error	Test: Sum of Squared Errors
7.8593	0.094867	0.308005	1664	832	278	0.140288	1	65.49308	0.117793	0.34321	556	277	554	0.155235	1	66.59307
7.8593	0.094867	0.308005	1664	832	278	0.140288	1	65.49308	0.117793	0.34321	556	277	554	0.155235	1	66.59307
1.8014	0.099039	0.314705	1664	832	278	0.147482	1	68.77056	0.123688	0.351693	556	277	554	0.162455	1	69.41467

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

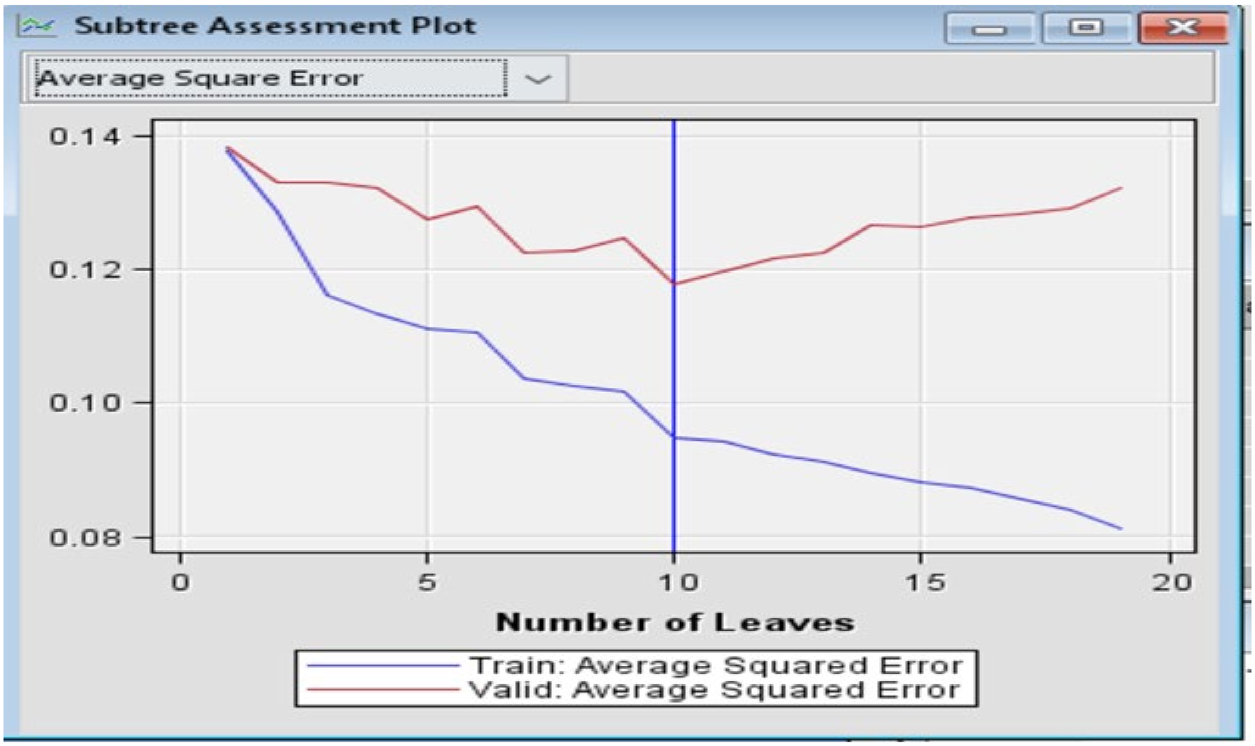
Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Decision Tree Model 2	0.14029	0.094867	0.11178	0.11779
	Tree3	Decision Tree Model 1	0.14029	0.094867	0.11178	0.11779
	Tree	Decision Tree	0.14748	0.099039	0.11899	0.12369

Decision Tree Model 1

Tree:



Sub – Assessment Plot – Misclassification Rate



From the above plot, we can see that misclassification rate increases after 10 leaves.

Interpretation and Accuracy of the best model

Classification Table – Decision Tree Model 1

Classification Table

Data Role=TRAIN Target Variable=Attrition Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	89.9204	97.5540	678	81.4904
YES	NO	10.0796	55.4745	76	9.1346
NO	YES	21.7949	2.4460	17	2.0433
YES	YES	78.2051	44.5255	61	7.3317

Data Role=VALIDATE Target Variable=Attrition Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
NO	NO	88.4462	95.6897	222	79.8561
YES	NO	11.5538	63.0435	29	10.4317
NO	YES	37.0370	4.3103	10	3.5971
YES	YES	62.9630	36.9565	17	6.1151

Event Classification Table

Data Role=TRAIN Target=Attrition Target Label=' '

False Negative	True Negative	False Positive	True Positive
76	678	17	61

Data Role=VALIDATE Target=Attrition Target Label=' '

False Negative	True Negative	False Positive	True Positive
29	222	10	17

Accuracy of the model can be determined by calculating accuracy, misclassification rate, true positive rate, false positive rate, true negative rate, precision and prevalence.

Misclassification Rate: Overall, how often is it wrong? (FP + FN)/TOTAL

Accuracy: True Positive (TP) + True Negative (TN) / Total. How often is the model correct?

True Positive Rate or Sensitivity or Recall (TP/Actual yes) – When it's actually yes, how often does it predict yes?

False Positive Rate (FPR) (FP/Actual no): When it's actually no, how often does it predict yes?

True Negative Rate (TNR) or Specificity (TN/Actual no): When it's actually no, how often does it predict no?

Precision: When it predicts yes, how often is it correct? (TP/predicted yes)

Prevalence (Actual Yes / TOTAL): How often does the yes condition occur in our sample?

Summary of the above results is as follows:

Accuracy	Misclassification Rate (FP+FN)	TPR	FPR	TNR	Precision	Prevalence
82.19%	0.104 (10.4%)	0.369 (36.9%)	0.4310 (4.310%)	0.956 (95.6%)	0.629 (62.9 %)	0.910 (9.1%)

Decision Tree Model -1 Accuracy summary

True Positive Rate (TPR) is low in our model which is 36.95% it means our model predicts 36.95% attrition rate correctly.

False Positive Rate (FPR) is 4.310% which means our model predicts 4.310% wrongly when the attrition rate is 'yes' but actually they are not.

True Negative Rate (TNR) is 95.6% which means attrition rate is 'yes' but model predicts correctly at 95.6%.

Variable Importance and Managerial Implication

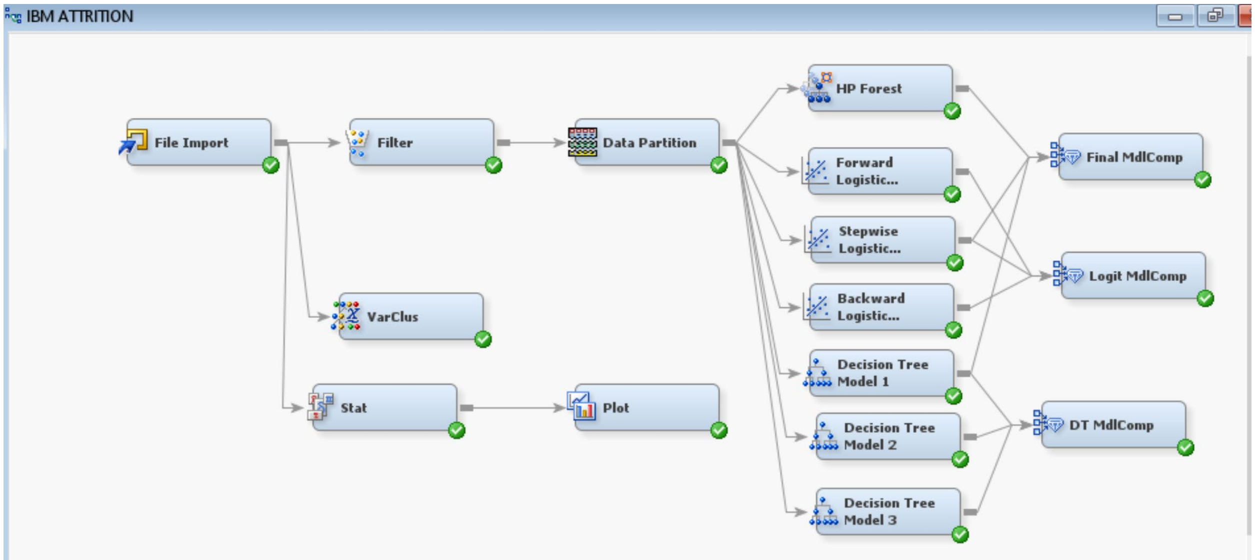
Variable Importance

Variable Name	Label	Numbe Splittin Rules
MonthlyInco...		
OverTime		
JobRole		
HourlyRate		
MaritalStatus		
DailyRate		
TotalWorkin...		
YearsInCur...		
Environme...		
StockOptio...		
EmployeeN...		
Education		
JobLevel		
JobSatisfac...		
NumComp...		
MonthlyRate		
StandardH...		
YearsWithC...		
DistanceFr...		
TrainingTi...		
VAR1	ï»¿Age	
WorkLifeBa...		
YearsAtCo...		
JobInvolve...		
YearsSince...		
Gender		
BusinessTr...		
Department		
EducationFi...		
Performanc...		
PercentSal...		
Relationshi...		

The above table shows the variables importance from the Decision Tree Model 1. It is evident that the most significant features are monthly income, overtime, job role, and hourly rate. Monthly income is the most important factor that affects employees leaving the organization. Overtime makes employees leave the organization and it clearly shows us that employees are not happy with the overtime rate. Job role is also an important factor, and it clearly says that employees are leaving the organization due to the unlikely role. Hourly rate is an important factor that makes employees leave the organization if they are not happy with the rate. These are the top four important features from our Decision Tree Model 1 for predicting the attrition rate. Organizations should consider these factors and plan to improve by making these changes to reduce the attrition rate.

Three different types of Decision Tree models were created in the project based on IBM Analytics HR dataset which can predict the attrition rate. Decision Tree Model 1 is the best model with 84.16% accuracy compared to all other models. The model has a high true positive rate and 70% AUC (Area Under Curve). This happened due to a highly imbalanced dataset target ratio. This is also the reason due to the fictional dataset. Trying with more training data, feature engineering and different algorithms can reduce the errors.

Final Model Comparison



Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg2	Stepwise Logistic Regression	0.12590	0.08214	0.10337	0.09846
	Tree3	Decision Tree Model 1	0.14029	0.09487	0.11178	0.11779
	HPDMForest	HP Forest	0.15468	0.10151	0.15264	0.11547

From the above figure, based on the validation misclassification rate, the model that minimizes the misclassification rate amongst Random Forest, Decision Trees and Logistic Regression is the Logistic Regression Model. Therefore, the best model we recommend to IBM management for deployment and prediction of new instances to minimize the cost effect of employee attrition is the Logistic Regression Model.

Conclusion

Employees are the backbone of any organization and the performance of an organization is dependent on retaining quality employees. Attrition is a problem that affects all businesses irrespective of their processes as it has an impact on productivity, profit, and time. Our focus on this report was to help the IBM human resources find resignation trends/numbers to understand the factors leading to employee attrition in the organization and provide a model to management that will predict attrition and control the rate. We looked specifically at the features, explored, prepared the dataset, and ran three different classification models and also cluster analysis. Binary logistic regression proved to be the best model after comparison using metrics such as misclassification rate. Logistic regression was also used to determine which variables contribute the most to attrition, and influence employee behavior. The following contributing factors were discovered after running our analysis: Business travel, Environment Satisfaction, Overtime, Job Satisfaction, Job Role, Stock Option level, Promotion and Number of Companies Worked. The model can be used from time to time to factor in new data from employees and discover more insights.

References

1. Gartner Survey Shows 29 Percent of Employees Witnessed At Least One Compliance Violation in The Last Two Years. (n.d.). Retrieved November 21, 2020, from https://www.gartner.com/en/newsroom/press-releases/2018-08-02-gartner-survey-shows-29-percent-of-employees-witnessed-at-least-one-compliance-violation-in-the-last-two-years.
2. “Here's How IBM Predicts 95% of Its Turnover Using Data.” LinkedIn Talent Blog, business.linkedin.com/talent-solutions/blog/artificial-intelligence/2019/IBM-predicts-95-\ percent-of-turnover-using-AI-and-data.
3. “Lack of Career Development Drives Employee Attrition.” Smarter with Gartner,www.gartner.com/smarterwithgartner/lack-of-career-development-drives-employee-attrition/.
4. Pavansubhash. “IBM HR Analytics Employee Attrition & Performance.” Kaggle, 31Mar.2017, www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.
5. Rohan’s Four - Rohan Jain, Ali Shahid. IBM HR Analytics Employee Attrition & Performance,inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/Jan uar y2018FBL/IBM_Attrition_VSS.html.
6. Zaal, T.M.E, and Steve Newton. Integrated Design and Engineering: as a Business Improvement Process. Maj Engineering Publishing, 2014.
7. https://towardsdatascience.com/people-analytics-with-attrition-predictions-12adcce9573f
8. https://www.clearpeaks.com/predicting-employee-attrition-with-machine-learning-using-knime/