Employee Attrition and Performance

Analysis on Employee Attrition

Employee Attrition and Performance

**Analysis on Employee Attrition**

# Final Report

Final Report

EXECUTIVE SUMMARY

**THE PROBLEM**

Employees are the most important part of an organization. Successful employees meet deadlines makes sales and build brands through positive customer interactions. In today’s competitive business environment, the impact of attrition on a business can be detrimental to both the bottom line and morale. Attrition can involve the loss of employees or the loss of customers, both employee turnover and failure to retain customers over time can challenge managers. Companies suffer from productivity losses and lost profits when there are large amounts of continuous churn in workforce. Top talent can be very difficult and expensive to replace. Employees Attrition is a major cost to an organization and predicting such attritions is the most important requirement of the Human Resources department in many organizations so that’s the problem we are addressing in this project)

**THE SOLUTION**

To reduce employee attrition, we must analyze what factors are contributing to the Employee attrition as it effects organization’s business in many ways like it increases the high training cost and the crucial business time of an organization. So, for that we will focus our analysis on these questions using the Employee Attrition Dataset taken from Kaggle. 1.We will first find the factors that lead to employee attrition? 2.Pridict Which Employees prone to leave the organization? 3.Which department is most contributing to the attrition rate of the company? 4.If attrition rate is related to productivity of an organization?

**MOTIVATION**

Each month in the US, 3 to 4.5 million employees quit their job. On-third of the employees quit after six months. This effects not only the operational cost but also revenue, productivity company culture, customer experience and more.

However, one survey also said 94% would stay at their current employer if they invested in their long-term goals. So, we wanted to explore what could make employee stay at their current companies.

Currently there are tools available in market to predict attrition of an employee. One such tool is **Peakon’s attrition prediction model**. It takes feedback from employees through predefined questionaries to get response on few measures like Engagement, Loyalty, Growth, Responsiveness, Tenure and Past resignations. As an output it gives a risk level metric as chance of attrition. But the model relies highly on the frequency of feedbacks, and it is more focused on loyalty of an employee. It does not consider the other environmental and socioeconomic factors, which are more closely related to the policies of company itself. We have considered all those factors in our study and model building. Thus, our model provides a 360-degree analysis, whereas the Peakon’s model unfairly provides an one sided view.

**BUSINESS QUESTION:**

**How to prevent employee attrition?**

We splitted the question in three parts and will find answers for the first two parts by applying

DM and BI techniques and models**.**

1.We will first find the factors that lead to employee attrition?

2.Pridict Which Employees prone to leave the organization?

3.What measures should be taken to reduce employee attrition?

**DATA OVERVIEW**

**Dataset Used for the project**

**Location:** US

**Source:** Kaggle

**Size of dataset:** 1470 Rows and 35 Columns

**Target Variable:** Attrition (YES or NO)

**Software Used:**

Microsoft Excel, SAS Enterprise Miner, SAS Studio, Jupyter Notebook

**DATA DICTIONERY**

**Data preprocessing:** The collected data is analyzed for their central tendency, variation, and shape. And filtered as well as converted to binary or numeric data to suit the statistical models.

**Data preparation and Understanding:**

The dataset consists of 35 variables and 1470 records. The variables are selected and transformed for our

study and analysis:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Definition** | **Transformed variable and variable type** |
| **Age** | **Age of Employee** | **Same as original dataset (Numerical)** |
| **Attrition** | **Attrition of Employee** | **Target variable (Categorical)** |
| **BusinessTravel** | **BusinessTravel of Employee** | **Travel\_Rarely (1,0), Travel\_Frequently (1,0) (Binary)** |
| **DailyRate** | **DailyRate of Employee** | **Same as original dataset (Numerical)** |
| **Department** | **Department of Employee** | **Department\_RND (1,0), Department\_Sales(1,0)(Binary)** |
| **DistanceFromHome** | **Employee DistanceFromHome** | **Same as original dataset (Numerical)** |
| **Education** | **Education of Employee** | **Same as original dataset (Numerical)** |
| **EducationField** | **EducationField of Employee** | **EducationField\_Life\_Science(1,0), EducationField\_Med(1,0),EducationField\_Marketing(1,0),EducationField\_HR(1,0),EducationField\_Tech(1,0)(Binary)** |
| **Employee Count** | **Employee Count of Employee** | **Same as original dataset (Rejected)** |
| **Employee Number** | **Employee Number of Employee** | **Same as original dataset (Rejected)** |
| **Environment Satisfaction** | **Environment Satisfaction**  **of Employee** | **Same as original dataset- (Categorical in 1 to 4 scale)** |
| **Gender** | **Gender of Employee** | **Gender\_female (1,0)(Female -1 Male-0)** |
| **Hourly Rate** | **Hourly Rate of Employee** | **Same as original dataset (Numerical)** |
| **JobInvolvement** | **JobInvolvement of Employee** | **Same as original dataset (Categorical in 1 to 4 scale)** |
| **JobLevel** | **JobLevel of Employee** | **Same as original dataset (Categorical in 1 to 5 scale)** |
| **JobRole** | **JobRole of Employee** | **Job\_role\_manager\_n\_above (1,0) : Manager and above- (Manager, Research Director, Manufacturing Director) . Else: (Sales Executive, Research Scientist, Lab technician, Healthcare Rep., Sales Rep., Human Resources)** |
| **JobSatisfaction** | **JobSatisfaction of Employee** | **Same as original dataset (Categorical in 1 to 4 scale)** |
| **MaritalStatus** | **MaritalStatus of Employee** | **MaritalStatus\_single (1,0), MaritalStatus\_married (1,0)** |
| **MonthlyIncome** | **MonthlyIncome of Employee** | **Same as original dataset (Numerical)** |
| **MonthlyRate** | **MonthlyRate of Employee** | **Same as original dataset (Numerical)** |
| **NumCompaniesWorked** | **NumCompaniesWorked**  **of Employee** | **Same as original dataset (Numerical)** |
| **Over18** | **If Employee is over 18 years** | **Yes/No -(1,0)** |
| **OverTime** | **If Over time applicable** | **Yes/No - (1,0)** |
| **PercentSalaryHike** | **PercentSalaryHike of Employee** | **Same as original dataset (Numerical)** |
| **PerformanceRating** | **PerformanceRating**  **of Employee** | **Same as original dataset (Categorical in 1 to 4 scale)** |
| **RelationshipSatisfaction** | **RelationshipSatisfaction**  **of Employee** | **Same as original dataset (Categorical in 1 to 4 scale)** |
| **StandardHours** | **StandardHours of Employee** | **Same as original dataset (Numerical)** |
| **StockOptionLevel** | **StockOptionLevel of Employee** | **Same as original dataset (Categorical in 0 to 4 scale)** |
| **TotalWorkingYears** | **TotalWorkingYears of Employee** | **Same as original dataset (Numerical)** |
| **TrainingTimesLastYear** | **TrainingTimesLastYear**  **of Employee** | **Same as original dataset (Numerical)** |
| **WorkLifeBalance** | **WorkLifeBalance of Employee** | **Same as original dataset (Categorical in 1 to 4 scale)** |
| **YearsAtCompany** | **YearsAtCompany of Employee** | **Same as original dataset (Numerical)** |
| **YearsInCurrentRole** | **YearsInCurrentRole**  **of Employee** | **Same as original dataset (Numerical)** |
| **YearsSinceLastPromotion** | **YearsSinceLastPromotion**  **of Employee** | **Same as original dataset (Numerical)** |
| **YearsWithCurrManager** | **YearsWithCurrManager**  **of Employee** | **Same as original dataset (Numerical)** |

This is the stage, where we explored the data by observing at its aspects and features of the dataset.

The basic measures of the dataset are central tendency, variation, and shape of the data. We check

these measures by measuring the mean, median and mode for central tendency; range, variance and

standard deviation for the variation; and skewness as well as kurtosis for measuring the shape.

We measure the mean as .

Variance is measured as and standard deviation is measured as .

As a measure of shape, skewness and kurtosis is measured as:



Distribution of attrition based on age of employees

Measures of Mean, Standard deviation, kurtosis and skewness of the variables

Attrition based on percentage of salary hike

**Chart, bar chart

Description automatically generated**

Chart, pie chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Scatterplot representation of the principal components show clearly identifiable

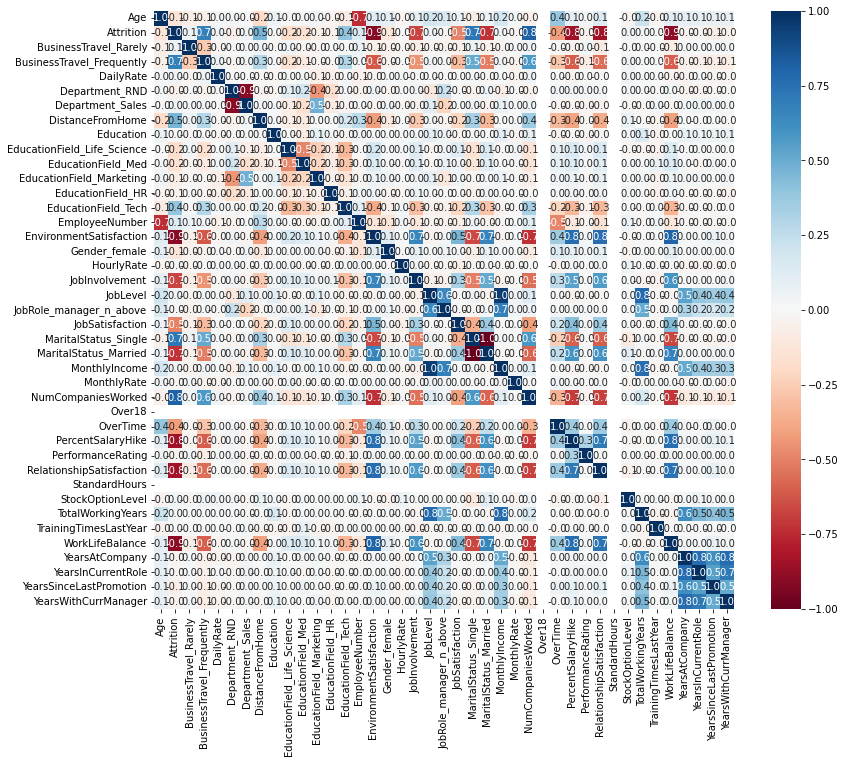
clusters of employees’ attributes when we group them by attrition. This visualization

confirms the assumption that some factors in our dataset are clearly influencing the

attrition of employees.

Graphical user interface

Description automatically generated with medium confidence



The above scatterplot matrix shows relationship between individual variables. Many interesting relationships can be established by looking at and analyzing the matrix plot. We can observe that- distance from home, percent salary hike, years at company have some relation with attrition of employee.

The heatmap in the right side showing the level of correlation of predictor variables. Blue color indicates positive correlation and red indicates negative correlation. Intensity of color is the scale.

**Decision Tree model:**

Within the data interpretation models, decision trees are the easiest to interpret

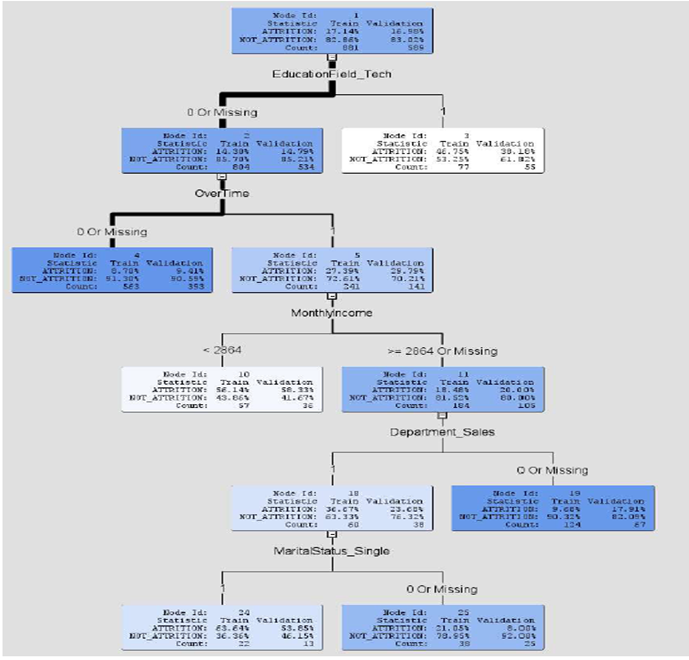
models. Decision tree model are also significantly transparent in comparison to other prediction

models. Trees work on the method of creating subgroups by separating and splitting records. Tree

splits are based on logical rules, which creates subgroups of more homogeneous nature. The

homogeneity is calculated by calculating Gini index or entropy for each node, till the subgroup node

reaches a threshold value.



Visual representation of decision tree model as SAS output

Using the decission tree model, we can build some rules to predict attrition:

1. IF(EducationField\_tech=0) AND IF(Overtime=0) Then Not\_Attrition=91%

2. IF(EducationField\_tech=1) Then Attrition=46.75%

3. IF(EducationField\_tech=0) AND IF(Overtime=1) AND IF(MonthlyIncome<2864) Then Attrition=56.14%

4. IF(EducationField\_tech=0) AND IF(Overtime=1) AND IF(MonthlyIncome>2864)

AND IF(Department\_sales=0) Then Not\_Attrition=90.3%

5. IF(EducationField\_tech=0) AND IF(Overtime=1) AND IF(MonthlyIncome>2864)

AND IF(Department\_sales=1) AND IF(Marital\_Status\_Single =1) Then Attrition = 63.6%

6. IF(EducationField\_tech=0) AND IF(Overtime=1) AND IF(MonthlyIncome>2864) AND

IF(Department\_sales=0) AND IF(Marital\_Status\_Single =0) Then Not\_Attrition = 78.9%

Table

Description automatically generatedGraphical user interface, application, table

Description automatically generated

Cumulative lift of the Decision tree model with training and validation data

Fit statistics for Decision tree model

**Performance of Decision tree model:**

**For training dataset:**

Accuracy of the model: 84.33%

Misclassification rate: 15.67%

Root mean squared error: 0.34

**For Validation dataset:**

Accuracy of the model: 84.21%

Misclassification rate: 15.79%

Root mean squared error: 0.35

Tr

**A screenshot of a computer

Description automatically generated with medium confidence**

Confusion matrix for Decision tree model

**Naïve Bayes classifier model:**

Using the Naïve Bayes formula we can calculate the probability of a record with a given set of predictor

values X1, X2, X3…Xn belonging to a perticulat class C1 among ‘m’ number of classes.

The formula can be written as:

**Diagram

Description automatically generated**

**A picture containing graphical user interface

Description automatically generatedTable

Description automatically generated with medium confidence**

Cumulative lift of the Naïve Bayes model with training and validation data

Fit statistics for Naïve Bayes classifier model

**Table

Description automatically generated**

**Performance of Naïve Bayes classifier model:**

**For training dataset:**

Accuracy of the model: 80.48%

Misclassification rate: 15.52%

Root mean squared error: 0.38

**For Validation dataset:**

Accuracy of the model: 79.46%

Misclassification rate: 20.54%

Root mean squared error: 0.39

Confusion matrix for Naïve Bayes classifier model

**Logistic regression model:**

In logistic regression, instead of using the f(x) = Y form as the outcome variable directly,

we use the logit function. The logit can be modelled as a linear function form of the predictors. As the

outcome variable of our dataset have a binary nature. Logistic regression suits the requirement. As this

model generates the probability of outcome variable as the output. The general form of logistic

regression is:

Where β0, β1,… βn are coefficients of predictor variables X1, X2, X3…..Xn. and p is the probability

P(Y=1)(*outcome variable*)

The output of logistic regression provides us with analysis of maximum likelihood estimates.

We can also identify the significant factors by looking into the p value of variables. For

the variables which can be regarded as significant for affecting attrition of employee are:

BusinessTravel\_Rarely, DistanceFromHome, EducationField\_Tech, EnvironmentSatisfaction,

JobInvolvement, NumCompaniesWorked , OverTime , PercentSalaryHike,TotalWorkingYears

and YearsSinceLastPromotion.

Table

Description automatically generated

Output of logistic regression- Analysis of maximum likelihood

**Chart, line chart

Description automatically generated**Table

Description automatically generated

Cumulative lift of the logistic regression model with training and validation data

Fit statistics of Logistic regression model

A screenshot of a computer

Description automatically generated with medium confidence

**Performance of Logistic Regression model:**

**For training dataset:**

Accuracy of the model: 85.92%

Misclassification rate: 14.08%

Root mean squared error: 0.33

**For Validation dataset:**

Accuracy of the model: 85.90%

Misclassification rate: 14.10%

Root mean squared error: 0.33

Confusion matrix of the Logistic regression model

**Comparison of models:**

Table

Description automatically generated

Summary report of confusion metrices of the three models

Chart, line chart

Description automatically generated

ROC curve: Sensitivity (true positive rate) vs.1- Specificity (1-true negative rate)

It shows that the logistic regression model has the highest accuracy in predicting true outcomes

Table

Description automatically generated

SAS output of model comparison node. Based on the parameters like misclassification rate and root mean squared error values, it suggests the logistic regression model as the best amongst the three models compared.

**Conclusion:**

Conducting experiments with three prediction models provides us with a scope to choose the best

model. To evaluate those models, we have used 60% of the whole dataset to train the models and 40%

of the dataset for validation purpose. We also compared the fit statistics, specifically the RMSE(*Root of*

*mean squared error*) to evaluate their relative performance.It is found that the Logistic regression

model provides highest accuracy and minimum misclassification rate with both training and validation

data. Using validation dataset,the logistic regression model have an accuracy of 85.90%, where the

decission tree and Naïve Bayes classifier have accuracy of 84.21% and 79.46% respectively. The RMSE

of logistic regression model is also lowest at 0.33, where the other two models have 0.35 and 0.39 as

RMSE respectively. Thereby, we can conclude that with the given variables, logistic regression model will

be best model for predicting attrition of an employee.

**Recommendations:**

Summing up our study based on three predictive models, we found that few factors are more

significant than others in predicting attrition of an employee. The significant factors are –

Business travel, Distance from home, technical education, Environment satisfaction, Job involvement,

Number of companies worked, Overtime facility, Percent of salary hike, Total working years, and

Years since last promotion. Also, the decision tree predicts that the employees who have technical

education and are working in sales department, have a higher attrition rate.

The company can alter their policies and resources to improve these factors as well as they can predict

probability of attrition of an employee by using the prescribed model.

**References:**

*Textbook:*

Data mining for business analytics

*By: Galit Shmueli, Peter. C. Bruce, Peter Gedeck, Nitin. R. Patel*

*External link:*

<https://support.peakon.com/hc/en-us/articles/360019702600-Understanding-Peakon-s-Attrition->