

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | SAR PROJECT REPORT | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | STUDENT’S NAME |  | | | |
|  | | | | Date—Course title—Teacher’s name |  | | | |
|  | | |  | | |  | | |

Contents

[SAR PROJECT REPORT 1](#_Toc144399608)

[STUDENT’S NAME 1](#_Toc144399609)

[Date 1](#_Toc144399610)

[— 1](#_Toc144399611)

[Course title 1](#_Toc144399612)

[— 1](#_Toc144399613)

[Teacher’s name 1](#_Toc144399614)

No table of figures entries found.

**Introduction**

Employee turnover, often known as churn rate, is a significant issue that firms in a variety of industries must deal with. Significant financial losses, poor productivity,and a negative effect on corporate culture can all be caused by high staff turnover.To address this issue proactively, we propose a project aimed at developing an advanced predictive model using R statistical techniques to forecast employee churn in our organization.

Employee turnover prediction involves using data analysis and predictive modeling to forecast which employees might leave a company. This is significant for: CostSavings: Minimizing expenses of hiring and training new employees. RetentionStrategies: Addressing issues to retain valuable employees. Succession Planning:Identifying successors for critical roles. Workforce Productivity: Planning for disruptions caused by turnover. Employee Engagement: Intervening to re-engage employees.

**Objective**

The primary objective of this project is to build a robust and accurate predictive model to forecast employee turnover. By utilizing historical employee data,including demographics and job-related factors, we aim to identify patterns and factors that contribute to churn, enabling us to take proactive measures for employee retention and engagement.

**Data Collection**

Dataset Overview The primary dataset has been downloaded from Kaggle. We are looking into other secondary datasets that can provide additional parameters/factors for making the above-mentioned prediction. The current dataset comprises of 16 columns, a brief description of which is as follows:

* ***stag*** – Tenure(in months)
* ***event***- Did the employee resign or not? (1/0)
* ***gender*** - Employee's gender (m/f)
* ***age***– Age in years, ranging from 18 to 58
* ***industry***- Industry in which the employee works
* ***profession***- The respondent's exact profession
* ***traffic***- From what pipeline the candidate came to the company
* ***coach***- Presence of a coach during probation
* ***head\_gender***- The supervisor's gender
* ***greywage***–Salary does not seem to the tax authorities(white/grey)
* ***way***-Medium of commute (by feet, by bus, etc.) -
* ***extraversion,independ (independent),selfcontrol,anxiety,novator (innovator)***

–Big 5 personality traits scored on a scale of 1 - 10

**Methodology**

To understand the dataset better, we did exploratory data analysis (EDA). As a part of EDA, we looked into the correlation between variables, comparative analysis of different industries, professions and also impact of factors like age, stag(tenure ),gender on the churn rate.

Post that we performed Logistic Regression as well as Linear Discriminant analysis on the dataset to generate the model which can classify the employees into two categories (whether they leave the organization of not). From both the models we got the similar accuracy levels (~63%) hence we decided to go with the LogisticRegression.

To check whether principle component analysis is required or not, we looked into the multicollinearity of the predictor variables and observed that they are not highly correlated. Hence we decided to finalize the Logistic Regression for the final model to predict the likelihood of employee leaving the organization.

**EDA**

Employee turnover between males and females across experience and age

It is seen that females have a higher attrition rate than males. The employee turnover for females is high when the experience is low.

A graph with red and blue dots

Description automatically generated

Employee turnover relation with Age of the employee

Below is the histogram plot between the age of an individual and the count of the individuals that turnover at that specific age. It is observed the turnover of individuals between 25 - 35 years of age is significantly high as compared to individuals who are above 40 years in age. As age increases the possibility of turnover decreases.

A graph of a graph

Description automatically generated

Employee turnover across profession

It is observed that except for IT and HR, for the rest of the professionals, the number of people who quit the job is higher than that of those who didn’t. All other professions have more than 50% attrition except for IT and HR. The highest attrition rates are observed in PR, Law and Engineer.

A graph with red and blue bars

Description automatically generated

Employee turnover across industries

It is observed that in almost half of the industry, the number of people who quit the job is higher than that of those who didn’t. The highest attrition rates are observed in Building, Agriculture and Banks.IT Industry is found to be a lucrative industry as it has the minimum attrition.

A graph with red and blue bars

Description automatically generated

Box plot for industry and traffic

The box plot's spread for a particular industry/traffic measure indicates the variability in tenure values for that industry/traffic. Greater spread implies higher variability in tenure. Industries with wider spreads likely have more diverse workforce traits, possibly causing different attrition trends. Industries with many outliers might undergo unusual attrition or have specific workforce segments with higher turnover.

A graph with lines and dots

Description automatically generated

A graph with black and white lines and dots

Description automatically generated

Correlation

The variables in this dataset do not show strong correlations between each other. There are no unusual or unexpected correlations in our matrix, which eliminates the possibility of many outliers or errors in data collection process.

A diagram of different colored squares

Description automatically generated

**Logistic Regression:**

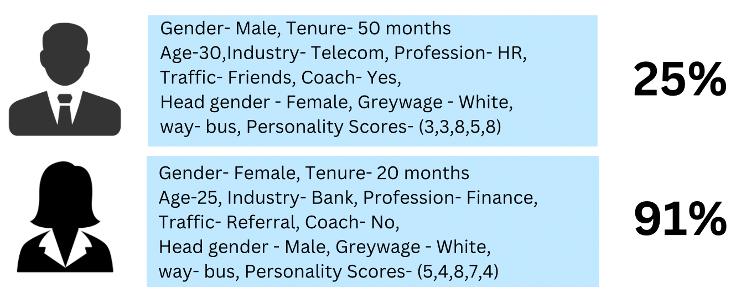
After completing an exploratory data analysis, a logistic regression model was run on a training data set (70% of total data), to classify attrition vs no attrition using all the parameters as input variables. Since the predicted variable (event) is a categorical binary variable (1/0), a logistic regression approach was more suited for the scenario. Other independent categorical variables were ‘dummified’ internally by R and a regression coefficient was estimated. Cut-off probability of classification was kept at default 0.5.

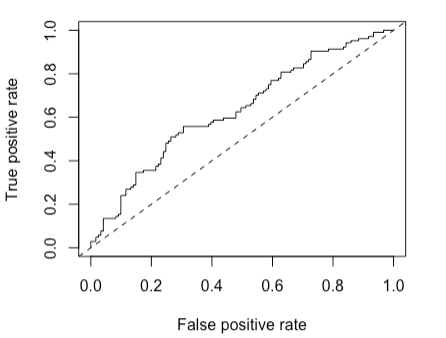
Probability was calculated as:

**p(X1,…,Xp)** = exp(β0+β1X1+⋯+βpXp)/(1+exp(β0+β1X1+⋯+βpXp))

*An excel sheet has been attached with all parameter’s coefficient and on changing input parameters, the model estimates probability that employee resigning (probability of attrition)*

***Two cases have been shown below with probability of resignation as estimated by model***

**

Post this the model accuracy and other parameters were estimated using the confusion matrix.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| Accuracy | 56.44% |
| Precision | 61.54% |
| Recall sensitivity | 52.46% |
| Specificity | 61.16% |
| AUC | 0.63 |

To understand whether variable shrinkage (ridge/lasso) or dimension reduction techniques like Principal Component Analysis is needed or not, VIF coefficients were inspected for all independent variables and all values were below 5, indicating no major collinearity. Nevertheless, logistic regression with PCA was conducted, but as expected it improved the model by just a small scale.

**Linear Discriminant Analysis**

As an alternative to Logistic Regression, a linear discriminant analysis was conducted on training data set using event as predicted variable and other variables as predictors and classification was conducted on test data keeping default cut off 0.5, post which model performance was estimated using different metrics like Logistics regression – the confusion matrix and thus the parameters were identical to that of logistic regression model.

**Logistic Regression & Linear Discriminant Analysis using PCA**

Logistic Regression models and LDA were re-run after running PCA on the 7 quantitative variables and reducing them to 4 components (explaining ~80% of variance), and predictors were taken as the 4 components and remaining categorical variables.

Following are the performance metrics using updated confusion matrix:

A graph with a line

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Value** | **Value (without PCA)** |
| Accuracy | 59.11% | 56.44% |
| Precision | 61.54% | 61.54% |
| Recall sensitivity | 55.17% | 52.46% |
| Specificity | 63.30% | 61.16% |
| AUC | 0.64 | 0.63 |

The metrics and ROCR curve improved to some extent, however the scale of improvement was quite less.

**Principal Component Analysis**

**Multiple Linear Regression**

As an additional analysis, a multiple linear regression was conducted keeping stag (tenure in months) as predicted variable and other variables (excluding event) as predictors. The model is meant to predict how many months an employee will stay before resigning. Since the regressand is a continuous variable, a simple linear regression was suitable for this scenario.

*The model has been attached in an excel sheet. An example case is being shown below:*

A blue rectangle with black text

Description automatically generated

Model Performance:

|  |  |
| --- | --- |
| Prediction MSE | Train MSE |
| 1043.2131 | 922.4698 |

**Conclusion  and Future Scope**

To summarize, we performed EDA to understand trend of resignation/attrition across different predictor variables and finally modeled a probability estimator to calculate attrition probability, given specific inputs with ~63% accuracy.

Future Scope:

Hybrid metric used for unbalanced class

Improving model accuracy by gathering more data

Adding more predictor variables (dimension reduction is applicable) to increase the efficacy and accuracy of model

Perform cluster analysis to create segments for better understanding

Industry applications:

Human resource management : Predicting possibility of attrition can help HRprofessionals design incentives and organizational structures effectively