Statistical Applications with R Poster Presentation

Predicting Employee Turnover/ Employee Churn Rate in Organisation:

Exploratory Analysis and Logistic Regression

Presented by: Aman Kumar, Praerit Agarwal, Rudra Banerjee, Shreyas Tupe, Aakash Verma, Prince Buddh, Samiksha Choudhary, Heeransh Singh

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: Cost Savings: Minimizing expenses of hiring and training new employees. Retention Strategies: 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.

Dataset and Methodology

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

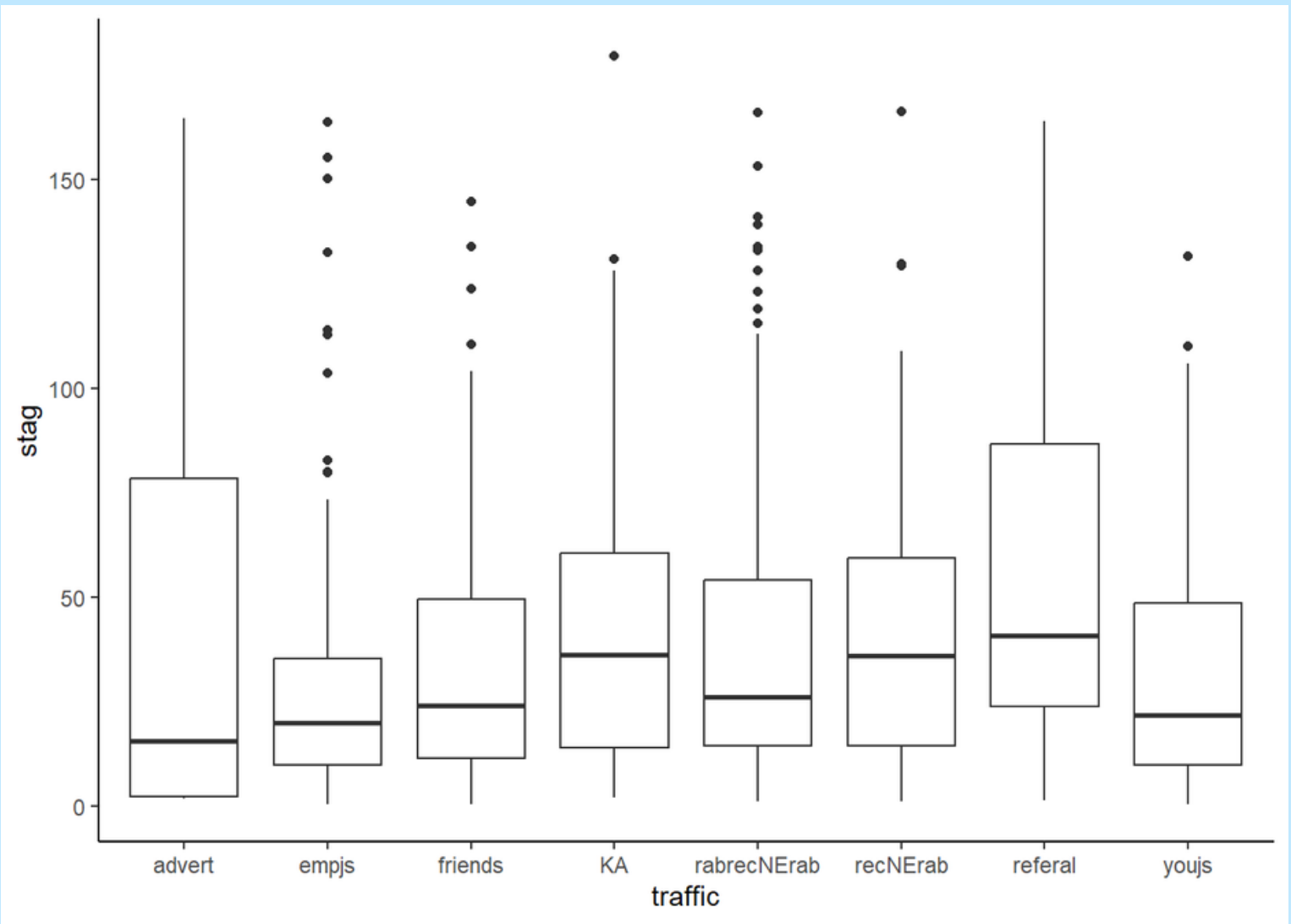
To understand the dataset better, we did exploratory data analysis (EDA). As a part of EDA, we looked into the correlation between variables, comparative analysis of different industries, professions and also impact of factors like age, stag(tenure), gender on the churn rate.

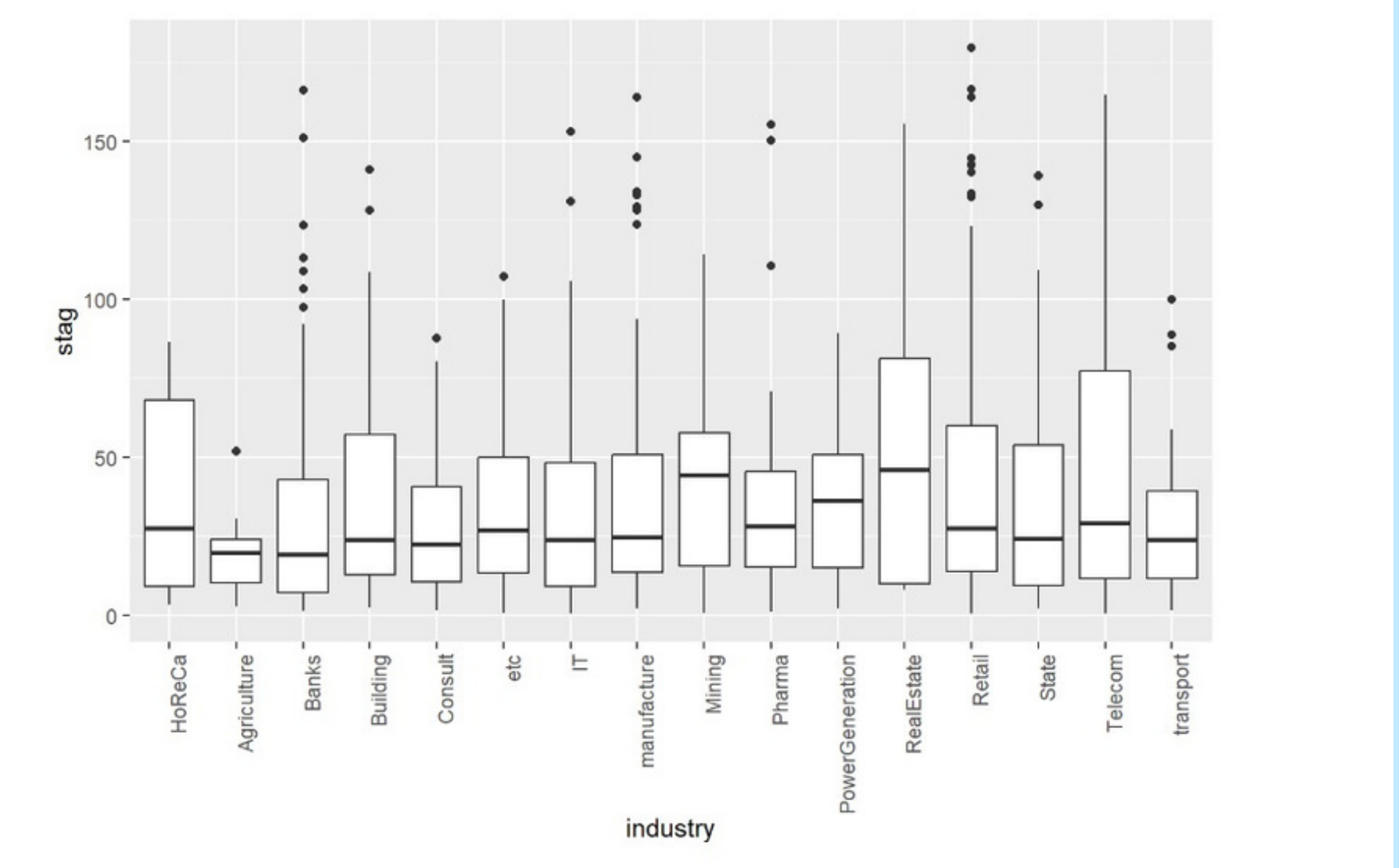
Post that we performed Logistic Regression as well as Linear Discriminant analysis on the dataset to generate the model which can classify the employees into two categories (whether they leave the organization of not). From both the models we got the similar accuracy levels (~63%) hence we decided to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

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 corelated. Hence we decided to finalize the Logistic Regression for the final model to predict the likelihood of employee leaving the organization.

Discussion on EDA

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.

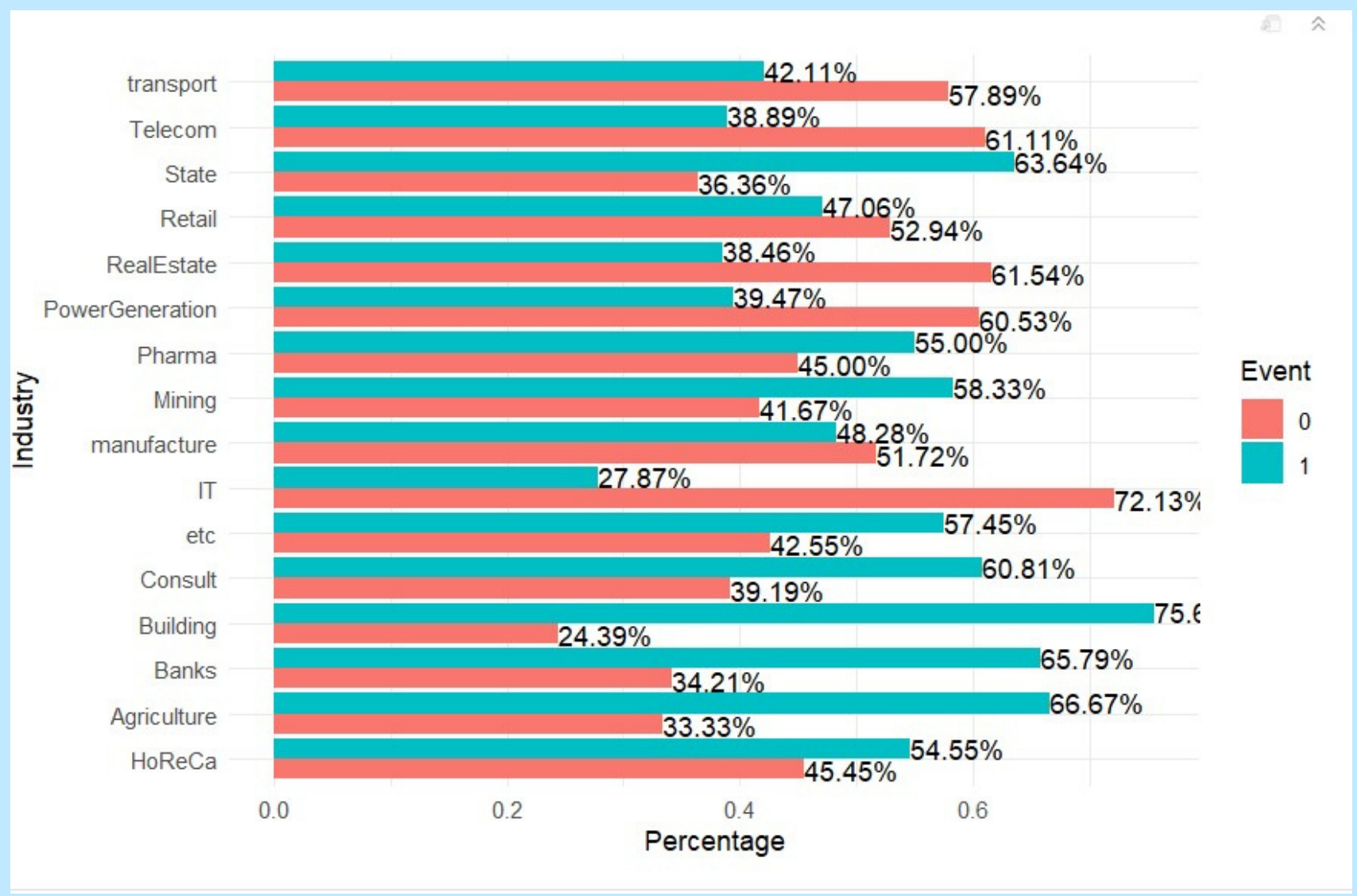


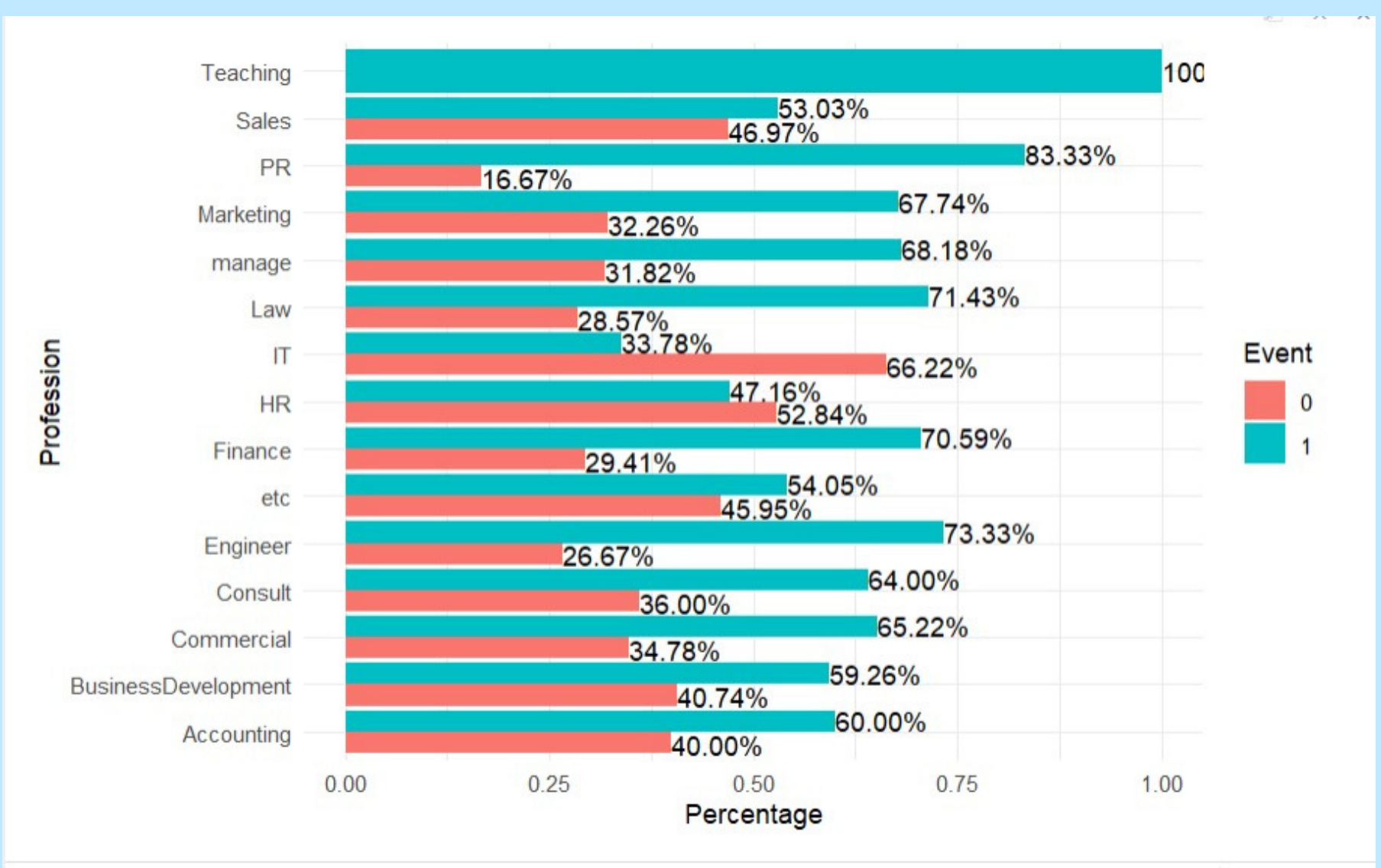


Discussion on EDA

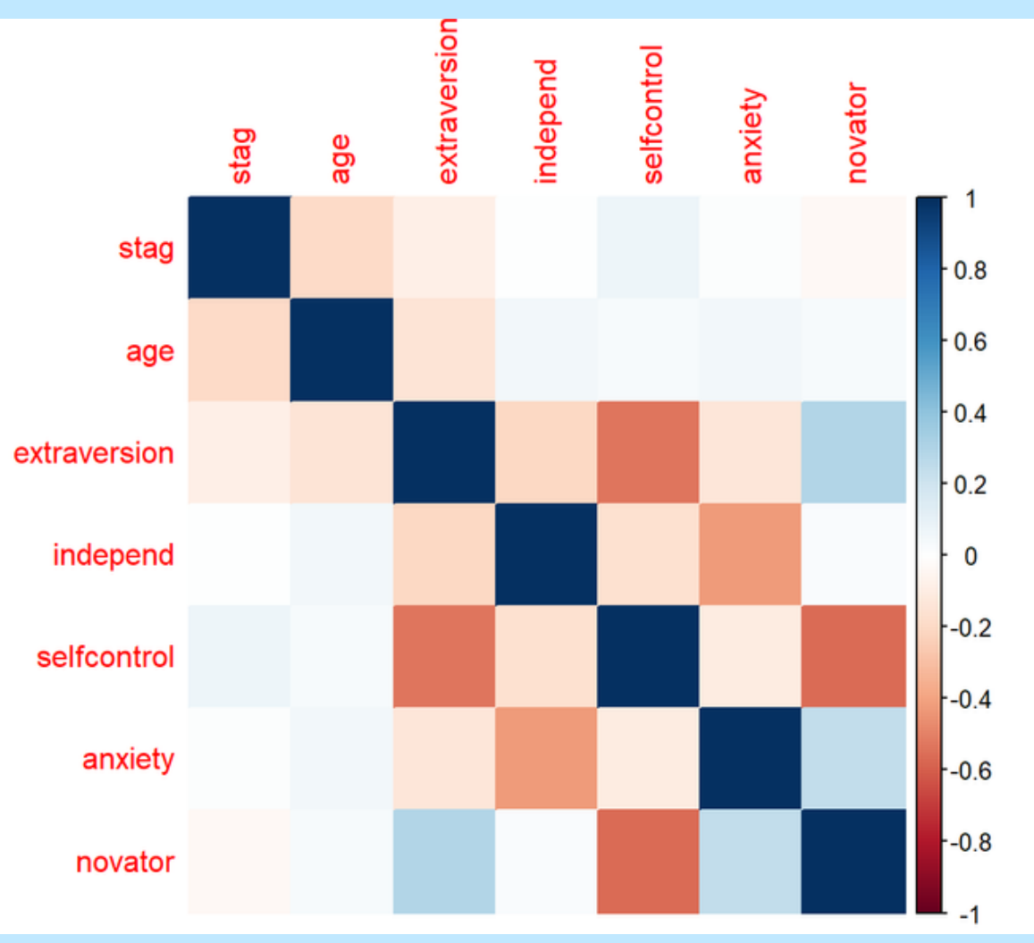
It is observed that except for IT and HR, for the rest of the professionals, the number of people who quit the job is higher than that of those who didn't. All other professions have more than 50% attrition except for IT and HR. The highest attrition rates are observed in PR, Law and Engineer

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.





Other variables in this dataset do not show strong correlation between each other which might contribute to the underperformance of the classifier. There are no unusual or unexpected correlations in our matrix, which eliminate the possibility of many outliers or errors in data collection process.



Discussion on Logistic Regression

We created a model which would predict the probability of an employee resigning given the input parameters. Examples are listed below

Gender- Male, Tenure- 50 months
Age-30,Industry- Telecom, Profession- HR,
Traffic- Friends, Coach- Yes,
Head gender - Female, Greywage - White,
way- bus, Personality Scores- (3,3,8,5,8)

25%

Gender- Female, Tenure- 20 months
Age-25, Industry- Bank, Profession- Finance,
Traffic- Referral, Coach- No,
Head gender - Male, Greywage - White,
way- bus, Personality Scores- (5,4,8,7,4)

91%

63% **Accuracy**
Overall Performance of the Model

62% **Precision**
How accurate the positive predictions are?

52% **Recall Sensitivity**
Coverage of actual positive sample

61% **Specificity**
Coverage of actual negative sample

56% **F1 Score**
Hybrid metric used for unbalanced class

Conclusion and Future Scope

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

Future Scope :

1. Improving model accuracy by gathering more data
2. Adding more predictor variables (dimension reduction is applicable) to increase the efficacy and accuracy of model
3. Perform cluster analsysis to create segments for better understanding

Industry applications:

1. Human resourse management : Predicting possibityof attrition can help HR professionals design incentives and organisational structures effectively