*Employee Turnover Analysis*

**1. Abstract**

The aim of this project was to explore employee turnover within a company. To yield results that held significance in the light of our main research goal, an Exploratory Data Analysis with data visualization plots and statistical tests along with multiple machine learning models were reviewed to extract a conclusion.

**2. Introduction**

Understanding employee turnover and factors that play into churn is a valuable asset to hiring organizations, staffing agencies, and management as it directly impacts productivity, cost, and company morale. By inspecting features play the biggest role in determining whether an employee will stay or leave the company, the organization can take proactive steps in preventing resignation for targeted personnel.

Through the use of a dataset sourced from Kaggle.com, the research investigated significant features’ influence in relation to employee resignation. Throughout the project, an Exploratory Data Analysis (EDA) was employed to uncover trends, visualize the dataset, and conduct statistical test which assessed attribute significance. Post EDA, multiple machine learning models were implemented to predict employee resignation which provided key indicators to turnover rates.

**3. Dataset**

The dataset used for this project was extracted from Kaggle.com and shared as a data card for a publication titled ‘HR Analytics’ that was updated 7 years ago by the username Giri Pujar. Pujar’s publication had earned a silver metal along with over 230 upvotes. Because of the public appreciation and over 200 independent projects this dataset was used in, it was deemed to be a safe and reliable csv file for the purpose of this project.

The original dataset consists of 15,000 rows and 10 features that are made up by a numeric column titled Satisfaction Level, a numeric column titled Last Evaluation, a numeric column titled Number of Projects, a numeric column titled Average Monthly Hours, a numeric column titled Time Spent with Company, a binary column titled Work Accident, a binary column titled Left (Resigned), a binary column titled Promotion within the Last 5 Years, a categorical column titled Department, and a categorical column titled Salary.

**4. Exploratory Data Analysis (EDA)**

**Figure 1**

*Statistical Summary of the Dataset*

A screenshot of a computer screen

Description automatically generated

As shown in Figure 1, the statistical summary of the dataset holds light to some valuable insights regarding the distribution of variables. For instance, satisfaction\_level with a mean of 0.6128 suggests a fairly balanced satisfaction among the employee records. Similarly, last\_evalution with a median of 0.72 implies that most workers were in good standing with the company. On average, most employees were involved with 4 projects and most had 8-hour workdays. While time\_spend\_company shows that tenure ranged from 2 to 10 years, many associates did not stay with their company for too long and a very small fraction of them had a work accident. About of quarter of the employee records show that they resigned and a very small percentage of them received a promotion. The two categorical features of the dataset show that most records were of sales associates and most people were assigned a medium threshold for their salary.

**Figure 2 & 3**

*Count of Employee Resignations & Distribution of Employee Tenure*

A graph showing a column of employees

Description automatically generated with medium confidenceA graph of a number of employees

Description automatically generated

Displayed above, Figure 2 visualizes the running number of associates who stayed or resigned from their respective company and Figure 3 is of the histogram of the distribution of workers’ accrued time with their company. Referring back to the time\_spend\_company feature displayed in Figure 1, we can draw the connection that most of the employees that make up the 3 to 6 bin in Figure 3 are the ones that also make up the ‘Resigned’ count in Figure 2.

**Figure 4 & 5**

*Proportion of Resignations by Promotions & Proportion of Resignations by Department*

A graph with red and blue squares

Description automatically generatedA graph of a company

Description automatically generated

As exhibited in Figure 4, a stark difference of employee resignations can be seen between those who have received promotions within the last 5 years and those who have not while Figure 5 presents a more uniform proportion of turnover rates. Examining Figure 4, it can be suspected that lack of career advancement opportunities can lead to lower satisfaction rates and resignations while career progression can positively increase the employee satisfaction levels and deter workers from seeking career advancements elsewhere. Inspecting Figure 5, it is apparent that employees in technical and cut-throat departments, such as Human Resources, Sales, and Technical Dept., are more likely to leave their job. This can be attributed to a high-pressure environment or lack of opportunities for growth. On the other hand, departments such as IT and Management can offer associates more stable job expectations and a larger room for career progression.

**Figure 6**

*Correlation Matrix of Features*

A chart with numbers and a red and green box

Description automatically generated with medium confidence

The correlation heatmap of the numeric variables displayed in Figure 6 grants us to examine strong and weak correlations between variables. For example, the relationship that has the strongest positive correlation is 0.42 between Number of Projects and Average Monthly Hours. It is quite simple to comprehend that the more projects an associate is tasked with, the more hours they would need to dedicate to their work. On the other hand, the relationship with the strongest negative correlation is -0.39 between Satisfaction Level and Left (Resigned). This behavior also makes sense for an employee to feel dissatisfaction towards their workplace and taking the action to leave said workplace. The stronger correlations are between variables the more important it would be to examine closely. For instance, there is no correlation between Work Accident, and Number of Projects therefore, it is not a priority to explore their relationship.

**Figure 7**

*Pearson's T-Test*

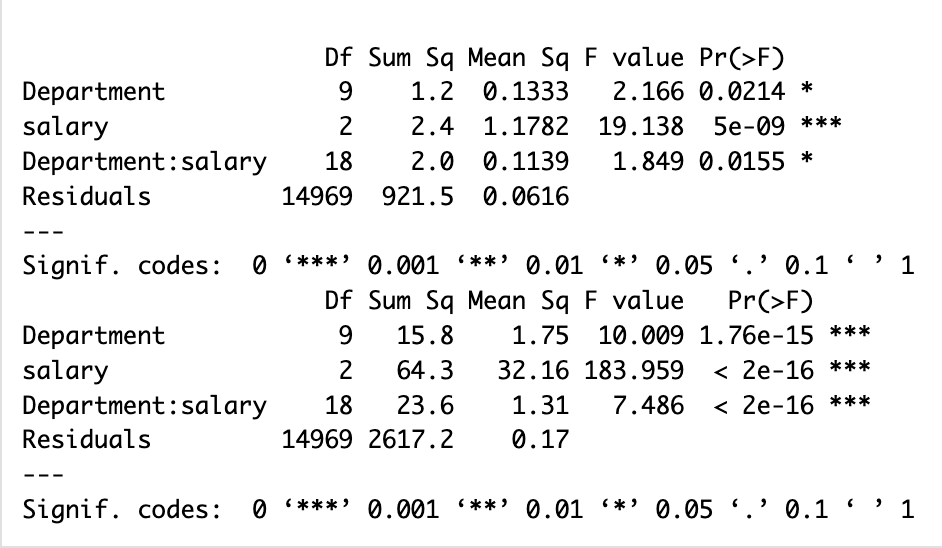
A table of numbers and symbols

Description automatically generated

The results displayed in Figure 7 reveal several features are significant in relation to employee resignation rates. Sorted by the P-Value from most significant to least, the satisfaction\_level variable displays a very strong significance to turnover rates therefore, employee satisfaction dropping from 0.67 to 0.44 is strongly associated with higher resignation rates. The Work\_accident variable displays a low P-Value as well, indicating that work accidents could be linked to higher resignations. The time\_spend\_company shows that employees who quit stayed with their respective company for an average of 3.88 years while those who stayed make up an average of 3.38. This finding can be linked to career plateauing or tenure burnout. Another statistically significant variable, promotion\_last\_5years, shows that employees who quit received far less promotions than the ones who stayed despite the average tenure of resigned associates being greater.

**Figure 8**

*Two-Way ANOVA: 1st Satisfaction Level 2nd Resignation*



As it appears in Figure 8, the two-way ANOVA tests done for satisfaction level and employee resignation contain remarkable significance for the categorical features of Department and Salary. The first ANOVA results display that both Department as well as Salary significantly impact associates’ satisfaction levels, with salary taking the lead at F = 19.138 and p = 5e-09. An even stronger significance of Department and Salary are determined for worker resignation. The highest level of significance on employees quitting is once again Salary at F = 183.959 and p < 2e-16. An additional remark to note is the dynamic between Salary and Departments with F = 7.486 and p < 2e-16 therefore, the interactions between turnover rates and employee salaries depend on departmental values.

**5. Models**

**Figure 9**

*Logistic Regression*

A graph with blue squares

Description automatically generated

The logistic regression model coefficient results displayed in Figure 9 examine the absolute coefficients of the most significant variables. The outcomes in Figure 9 reflect our previous findings regarding the sheer influence of satisfaction levels on employees choosing to stay at a job or not. To evaluate the supervised machine learning model, logistic regression, a confusion matrix comes in handy. With 10678 true negatives and 1074 true positives, the applied logistic regression model is assigned an accuracy rate of 0.7835.

**Figure 10**

*Random Forest*

A screenshot of a computer

Description automatically generated

Performing the best out of the implemented machine learning models, Figure 10 displays the evaluation matrix of the supervised machine learning model, Random Forest (RF). With an accuracy score of 0.894 and sensitivity of 0.9755, (RF) was the best suited prediction modelling method for the employee turnover dataset. Therefore, in the case of an organization’s interest in predicting current employee’s risk of resignation, the (RF) model trained after the HR Analytics would have the best chance at targeting employees at risk and implementing preventative measures to ensure company loyalty.

**Figure 11**

*Support Vector Machines (SVM)*

A screenshot of a computer

Description automatically generated

Scoring slightly better than the No Information Rate, the supervised machine learning model, Support Vector Machines (SVM), is displayed in Figure 11. With an accuracy score of 0.7786, the SVM’s Kappa statistic of 0.2378 indicates a fair agreement between predictions and actual outcomes. Though the performance of the model is not ideal, it is also not unacceptable. The evaluation metrics displayed in Figure 11 suggest that for the application on the HR Analytics dataset, the SVM model might require further examination and optimization.

**6. Conclusion & Summary of learning**

In short, this project underscores the sway of job satisfaction, fair monetary compensation, and room for professional growth in respect to employee turnover. By improving employee morale, especially among high stress departments, organizations can retain their workforce and by using Random Forests can target associates at risk.

My takeaways from this project has consisted of going through many trial and errors from finding the right dataset to figuring out the right syntax for the machine learning models but, overall I found this analysis to be a very intriguing exploration for me and there had been moments where I was surprised by the outcomes of my EDA and models.