Employee Attrition Analysis Report

Objective

The objective is to develop a logistic regression model to predict employee retention based on demographic details, job satisfaction scores, performance metrics and tenure. By leveraging these insights, the company can refine its retention strategies, foster a more supportive work environment and enhance workforce stability and satisfaction.

Overall approach of building logistic regression model is listed below:

1) Data understanding

The data has 24 Columns and 74610 Rows which includes details such as demographic details, job satisfaction scores, performance metrics, and tenure.

2) Data Cleaning

- Checked for the missing values:

'Company Tenure' and 'Distance from Home' columns had missing value percentage of 2.56 and 3.23 respectively hence imputed it with their median values.

- Checked for redundant columns:

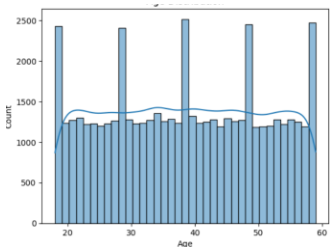
No such redundant columns are found in the dataset but Employee_ID is not needed for modelling, hence dropped it.

3) Train -Validation split

Split the entire dataset as validation and train data. We build the model using train data and tested on validation data. Target variable is Attrition which has info whether employee will stayed or left.

4) EDA on training data

Univariate Analysis: Age distribution: based on the plot, most of employees are around age 40



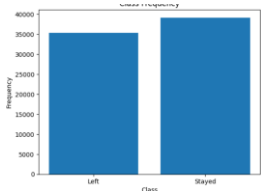
Please verify jupyter notebook for plots of below variables.

- Years at company: based on plot most of the employees have 30 years of experience
- Monthly income: most of the employees have income less than 20000
- Number of promotions: most of the employees have either 1,2 or no promotions
- No dependents: most of employees don't have dependents

correlation analysis: years at company and age have correlation of 0.5.Please verify jupyter notebook for heatmap plotted.

Check class balance:

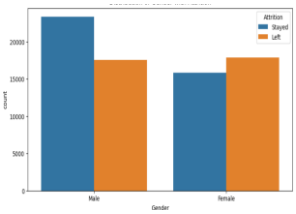
Based on the below plot on target variable-Attrition.We can infer that most of the employees have stayed.



Bivariate Analysis:

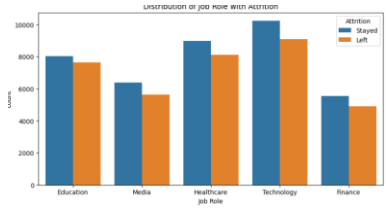
Gender with Attrition:

we can infer from below plot that most the male employees have stayed**.**



Job role with Attrition:

we can infer from below plot that most the employees in Technology job role are likely to stay



Please verify jupyter notebook for plots of below variables.

- Work-life balance with Attrition: employees who feel there is good work-life balance are likely to stay and Fair are likely to leave.
- Job stisfaction with attrition: employees with high job satisfaction are likely to stay and low are likely to leave.
- Perforrmance rating with attrition: employees with average Perforrmance rating are likely to stay and low are likely to leave.
- Overtime with attrition: employees with no overtime are likely to stay and has overtime are likely to leave.
- Education with attrition: employees with bachelors degree are likely to stay.
- Marital status with attrition: married employees are likely to stay and singles are likely to leave.
- Job level with attrition: employees with entry level job are most likely to leave the company and mid level employees are likely to stay.
- Companhy size wit attrition: medium company size people will stay.
- Remote work with attrition:No remote work people are likely to leave the company.
- Employees with with good company recognition are likely to stay.
- If company reputation are poor or fair employees are likey to leave.

5) EDA on Validation data: performed univariate, bivarite, correlation,check balance analysis on validagion data.please verify jupyter notebook.

6) Feature engineering:

Created dummy variables on categorical variables on both train and validation data.Applied feature scaling to the numeric columns.

7) Model Building:

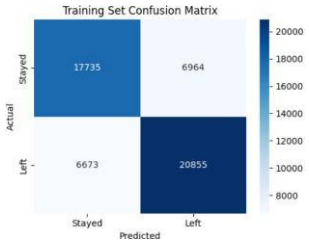
Model Interpretation: Used RFE to select 15 columns and built logictic regression model and got statistical aspects. P-values for all features are 0 indicating all features are more significant and VIFs are less for all features. Hence proceeded to next step to perform predictions.

Model prediction:

- accuracy of the model based on the predictions made on the training set is 0.739 ~73.9%
- confusion matrix based on the predictions made on the training set

[[17735 6964]
[6673 20855]]

Indicates that 6673 employess actually left but predicted as stayed and 6964 actually stayed but wrongly predicted as left. 17735 actually stayed and predicted as stayed. 20855 actually left and predicted the same



Other metrics calculated for train data are:

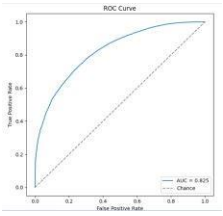
Sensitivity (Recall): 0.758, Specificity: 0.718, Precision: 0.750, Recall: 0.758

Sensitivity indicating 75.8% of employees who left their job and were correctly predicted to leave their job

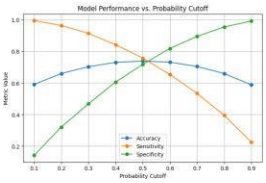
Specificity indication 71.8% of employees who stayed with the company who were correctly predicted to stay with the company.

Plotted ROC curve and checked AUC:

Below plot has ROC curve with 45 deg diagonal and AUC as 0.825 indicating good model.



Find the optimal cutoff: Based on below plot for accuracy, sensitivity, specificity at different probability cutoffs we can infer all intersect at nearly 0.5,Hence optimal cutoff is considered as 0.5.



final prediction based on the optimal cutoff:

Final training accuracy @ cutoff 0.50: 0.739 ~73.9%

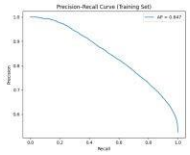
Final Confusion Matrix:

[[17735 6964]

[6673 20855]]

Sensitivity (Recall): 0.758, Specificity: 0.718, Precision: 0.750, Recall: 0.758

Check optimal cutoff value by plotting precision-recall curve: from below plot AP is 0.847 which is higher value near to 1 indicating better performance.



8) Model Evaluation:

accuracy of the model based on the predictions made on the validation set is 0.737~73.7 which is almost near to accuracy value od train set.

Confusiton matrix of validation set:

[[7649 3071] indicates 2810 employess actually left but predicted as stayed and

[2810 8853]] 3071 actually stayed but wrongly predicted as left

Other metrics calculated for validation data are:

Sensitivity (Recall): 0.759

Specificity: 0.714

Precision: 0.742

Recall: 0.759

- Sensitivity indicating 75.9% of employees who left their job and were correctly predicted to leave their job
- Specificity indication 71.4% of employees who stayed with the company who were correctly predicted to stay with the company.

All metrics are nearly same for both train and validation dataset. Hence, we can infer that we have built a good model.

Overall we can conclude that we have built a decent model to predict employee retention (whether employee stay or leave the company). HR department can use the analysis made while building model and also use this model to make better strategies to retain employees in their company.

Key takeaways:

- Employees working overtime and lacking remote work privileges are more likely to leave.
- Improving work-life balance, recognition programs, and career growth opportunities may reduce attrition.
- Demographics such as gender and marital status also influence employee stability and retention.

Author:
Tiyasa Mukherjee
Akshatha B