



Machine Learning

Course-End Project Problem Statement

Course-End Project: Employee Turnover Analytics

Project Statement:

Portobello Tech is an app innovator that has devised an intelligent way of predicting employee turnover within the company. It periodically evaluates employees' work details including the number of projects they worked upon, average monthly working hours, time spent in the company, promotions in the last 5 years, and salary level.

Data from prior evaluations show the employee's satisfaction at the workplace. The data could be used to identify patterns in work style and their interest to continue to work in the company.

The HR Department owns the data and uses it to predict employee turnover. Employee turnover refers to the total number of workers who leave a company over a certain time period.

As the ML Developer assigned to the HR Department, you have been asked to create ML Programs to

1. Perform data quality check by checking for missing values if any.
2. Understand what factors contributed most to employee turnover by EDA.
3. Perform clustering of Employees who left based on their satisfaction and evaluation.
4. Handle the left Class Imbalance using SMOTE technique.
5. Perform k-fold cross-validation model training and evaluate performance.
6. Identify the best model and justify the evaluation metrics used.
7. Suggest various retention strategies for targeted employees.

Data will be modified from:

<https://www.kaggle.com/liujiaqi/hr-comma-sepcsv>

Column Name	Description
satisfaction_level	satisfaction level at the job of an employee

last_evaluation	Rating between 0 to 1, received by an employee at his last evaluation
number_project	Number of projects, an employee involved in
average_monthly_hours	Average number of hours in a month, spent by an employee at office
time_spend_company	Number of years spent in the company
Work_accident	0 - no accident during employee stay, 1 - accident during employee stay
left	0 indicates employee stays in the company, 1 indicates - employee left the company
promotion_last_5years	Number of promotions in his stay
Department	Department, an employee belongs to
salary	Salary in USD

Perform the following steps:

1. Perform data quality check by checking for missing values if any.
2. Understand what factors contributed most to employee turnover by EDA.
 - 2.1. Draw a heatmap of the Correlation Matrix between all numerical features/columns in the data.
 - 2.2. Draw the distribution plot of
 - Employee Satisfaction (use column satisfaction_level)
 - Employee Evaluation (use column last_evaluation)
 - Employee Average Monthly Hours (use column average_monthly_hours)
 - 2.3. Draw the bar plot of Employee Project Count of both employees who left and who stayed in the organization (use column number_project and hue column left) and give your inferences from the plot.
3. Perform clustering of Employees who left based on their satisfaction and evaluation.
 - 3.1. Choose columns satisfaction_level, last_evaluation and left.

-
- 3.2. Do KMeans clustering of employees who left the company into 3 clusters.
 - 3.3. Based on the satisfaction and evaluation factors, give your thoughts on the employee clusters.
 4. Handle the left Class Imbalance using SMOTE technique.
 - 4.1. Pre-Process the data by converting categorical columns to numerical columns by
 - Separating categorical variables and numeric variables.
 - Applying get_dummies() to the categorical variables.
 - Combining categorical variables and numeric variables.
 - 4.2. Do the stratified split of the dataset to train and test in the ratio 80:20 with random_state=123.
 - 4.3. Upsample the train dataset using SMOTE technique from the imblearn module.
 5. Perform 5-Fold cross-validation model training and evaluate performance.
 - 5.1. Train a Logistic Regression model and apply a 5-Fold CV and plot the classification report.
 - 5.2. Train a Random Forest Classifier model and apply the 5-Fold CV and plot the classification report.
 - 5.3. Train a Gradient Boosting Classifier model and apply the 5-Fold CV and plot the classification report.
 6. Identify the best model and justify the evaluation metrics used.
 - 6.1. Find the ROC/AUC for each model and plot the ROC curve.
 - 6.2. Find the confusion matrix for each of the models.
 - 6.3. From the confusion matrix, explain which metric needs to be used- Recall or Precision?
 7. Suggest various retention strategies for targeted employees.
 - 7.1. Using the best model, predict the probability of employee turnover in the test data.
 - 7.2. Based on the below probability score range, categorize the employees into four zones and suggest your thoughts on the retention strategies for each zone.
 - Safe Zone (Green) (Score < 20%)
 - Low Risk Zone (Yellow) (20% < Score < 60%)



- Medium Risk Zone (Orange) ($60\% < \text{Score} < 90\%$)
- High Risk Zone (Red) ($\text{Score} > 90\%$).