Employee Attrition Prediction and Analysis of Google

Project Goal

Employee attrition prediction models will help Google's HR team to proactively identify and mitigate risks in terms of employee turnover. Through the analysis of employees' historical data and identification of the key factors driving attrition, the project aims at smartening employee retention strategies and increasing the satisfaction levels.

Workflow Overview

Data Collection and Preparation

Data Source: HR dataset containing information on employee demographics, job details, performance metrics, and attrition records.

Data Points:

Demographics: Age, Gender, Marital Status.

Job Details: Job Role, Department, Years with the Company, Current Salary.

Performance Metrics: Job Satisfaction Score, Performance Rating, Training History, and Work-Life Balance Score.

Attrition Details: Exit Reason, Status (Active/Terminated), Resignation Status, and Date of Leaving.

Data Cleaning :

Missing values in "Date of Leaving" were imputed with place holders for active employees.

Encoding done on categorical data like Gender, Department

Scaling of numerical values to enable training of the model.

Exploratory Data Analysis

Tools/Software: Microsoft Excel and Power BI.

Excel:

All the trends and anomalies sensed through data validation and conditional formatting

Built pivot tables and charts to see department-wise attrition and distribution of tenures

Mean tenure, median salary calculated for getting a sense of major patterns descriptive statistical measure

Power BI

Imported cleaned data to get interactive visualizations.

High employee attrition rates for departments and jobs

Low job satisfaction scores of employees being between 1-2 and work-life balance score of employees between 1-2 had higher chances of leaving the organization.

Attrition mostly occurred in the early tenure below 5 years

Visualisations Developed:

Attrition Rates by Department (Excel & Power BI)

Histogram of Job Satisfaction Scores (Excel & Power BI)

Attrition rate by tenure (Excel & Power BI)

Feature Engineering

Some Features extracted

Tenure-in days : It is the difference between leaving and joining dates.

Salary Progression: Staring vs. Current Salary

Training History Completion Rate: Translated from "Training History."

Correlation Analysis

First attempted in Excel using correlation matrices and then in Python to identify the most highly correlated with attrition.

Model Building

Data Split; Trained on 80% of data and tested on 20%

Algorithm Selection

Logistic Regression was a baseline prediction

Used Gradient Boosting (XGBoost) for greater accuracy

Model Summary

Logistic Regression: 78% accuracy, F1 score of 0.74

XGBoost: 85% accuracy, F1 score of 0.81

Feature Importance:

Job Satisfaction Score and Work-Life Balance Score emerged as the strongest predictors.

Model Interpretability

Tools Used: SHAP (SHapley Additive exPlanations).

Outcome:

Attrition had been mainly affected by the low job satisfaction scores.

Employees working in stress-generating jobs such as IT and Management were more at risk

Reporting

Excel Report:

Published aggregate sheets on key metrics like average tenure, attrition rates, rating of performance.

Grpahs and charts are added to show attrition and satisfaction scores vs departments.

Power BI Report :

An interactive report that summarized trends and key findings.

Incorporated the attrition rates at the department level, tenure analysis, and impact on the satisfaction score.

Recommendations

Revamped work-life balance and satisfaction in high-turnover roles

Focused training and development for employees who possess less than satisfactory scores

Deployment and Automation

Model Deployment:

Scheduled Python scripts to periodically retrain and predict.

Attrition predictions were fed into HR systems to offer proactive retention measures

Automation:

SQL and Python scripts automatically updating data and generating reports

Tools and Technologies

Excel: Validations, descriptive statistics, pivot tables, and charts

SQL: Cleaning of data, feature engineering, and preprocessing

Power BI: Exploratory analysis and reporting

Python: Development of a machine learning model using Scikit-Learn and XGBoost.

Results

Model Accuracy : 85% was achieved by the XGBoost model

Insights

Job dissatisfaction, poor work-life balance, and jobs characterized by high stress are the three main reasons for attrition

Recommended Interventions

Career development and better work-life policies can be recommended as the interventions

Impact on Organization

The organization can bring down the turnover rate through proactive retention strategies and help reduce costs due to it.

Targeted initiatives by HR for increased satisfaction and engagement.

This report will take into account all views as can be gathered from the Excel and Power BI analysis with the machine learning component. Do you want further visualization or format for presentation to stakeholders?