HR Analytics - Predict Employee Attrition

Introduction

Employee attrition, or the rate at which people leave an organization, has a big impact on both company culture and costs. When employees leave, companies spend significant time and money recruiting and training replacements, while also facing potential drops in productivity and morale. To address this challenge, predictive modeling offers a way to identify which employees are most likely to leave. By spotting early warning signs, organizations can take proactive steps to improve retention and keep their teams engaged.

In this project, we built a machine learning model that predicts the likelihood of an employee leaving their job. To better understand what drives these predictions, we used **SHAP** (**SHapley Additive exPlanations**), a powerful tool that explains how each feature—like overtime hours, job satisfaction, or years at the company—contributes to the model's decisions. This approach helps transform complex data into clear insights that can guide smarter, more effective retention strategies.

Abstract

This project involves building a logistic regression model to predict the likelihood of employee attrition based on various features such as overtime status, job satisfaction, tenure, and marital status. After training the model, SHAP values are computed to provide interpretability. SHAP summary plots help visualize the impact of each feature on the model output, enabling data-driven insights into attrition drivers.

Tools Used

- **Python**
- **Pandas** – Data preprocessing
- **Matplotlib/Seaborn** – Visualization
- **Scikit-learn** – For logistic regression modeling
- **SHAP** – For model interpretability

Steps Involved in Building this Project

1. **Data Preprocessing** - Collected and cleaned HR data containing features such as: Age, Overtime, Marital status, Job involvement and satisfaction, Years with the current manager and Encoded categorical variables (e.g., `OverTime_Yes`, `MaritalStatus_Single`).

2. **Model Training** - A logistic regression model was trained to classify whether an employee is likely to leave or stay.
3. **SHAP Explainer** - Corrected the error by using `shap.LinearExplainer`, suitable for linear models like logistic regression.

4. **Computing SHAP Values**
    - SHAP values were computed to quantify the contribution of each feature to predictions.
    - Each SHAP value indicates how much a feature shifts the prediction away from the baseline.

5. **Visualization**
    - Generated SHAP summary plots to rank feature importance.
    - Colors represented feature values (blue = low, red = high), and positions along the x-axis showed positive or negative impact on attrition risk.
    - Key findings from the plots:
        - High overtime and being single increased attrition probability.
        - Greater satisfaction and tenure reduced attrition risk.
        - Younger employees were slightly more prone to leave.

Conclusion

This project demonstrated how logistic regression combined with SHAP explanations can effectively predict and interpret employee attrition risk. By understanding the most impactful factors, organizations can design better retention strategies—such as improving work-life balance, reducing excessive overtime, and providing targeted support to at-risk employees. SHAP visualizations were instrumental in translating the model's decisions into actionable business insights.
