**Employee Turnover Analysis and Retention Strategies**

**Introduction**

Portobello Tech, an app innovator, has devised an intelligent way of predicting employee turnover within the company. The HR Department periodically evaluates employees' work details, including the number of projects they worked on, average monthly working hours, time spent in the company, promotions in the last 5 years, and salary level. This data is used to identify patterns in work style and their interest in continuing to work in the company. As the ML Developer assigned to the HR Department, the objective is to create machine learning programs to analyze and predict employee turnover and suggest retention strategies.

**Data Quality Check**

The dataset was first checked for missing values and data types. It was found that there were no missing values, and the data types were appropriate for analysis. Categorical variables such as 'department' and 'salary' were converted to dummy variables for further analysis.

**Exploratory Data Analysis (EDA)**

**Correlation Analysis**

A correlation matrix was created to understand the relationships between numerical features. Key highlights include:

* A moderate positive correlation (0.42) between the number of projects and average monthly hours, indicating that more projects result in higher workload.
* A moderate negative correlation (-0.39) between satisfaction level and turnover, suggesting that lower satisfaction levels are associated with higher turnover rates.

**Distribution Analysis**

The distribution of employee satisfaction, evaluation scores, and average monthly hours was plotted. Key observations include:

* A significant portion of employees rated high satisfaction, but there was also a notable peak of extremely low satisfaction.
* The distribution of evaluation scores showed a bimodal pattern, indicating two distinct groups of employees.
* Average monthly hours also showed two main groups: one working around 150 hours and the other working over 250 hours monthly.

**Project Count Analysis**

A bar plot of the number of projects versus turnover revealed three main behaviors:

* Employees with 2 projects had a high turnover rate, possibly feeling underutilized.
* The optimal number of projects was 3 to 4, where turnover was minimal.
* Employees with 5 or more projects had a significantly higher turnover rate, indicating excessive workload.

**Clustering Analysis**

Using unsupervised learning (KMeans clustering), employees who left the company were clustered based on their satisfaction level and last evaluation score. Three clusters were identified:

1. **High Evaluation, Low Satisfaction (Yellow Cluster):**
   * Employees with high evaluation scores but low satisfaction levels.
   * Implication: High performance does not necessarily equate to high retention if satisfaction is low.
2. **Low Evaluation, Low Satisfaction (Purple Cluster):**
   * Employees with both low evaluation scores and low satisfaction levels.
   * Implication: Poor performance coupled with dissatisfaction leads to higher turnover.
3. **High Evaluation, High Satisfaction (Teal Cluster):**
   * Employees with high evaluation scores and high satisfaction levels.
   * Implication: External factors such as better opportunities or personal reasons might influence their decision to leave.

**Handling Class Imbalance**

The dataset exhibited class imbalance, with fewer employees leaving the company compared to those staying. The SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the classes before model training.

**Model Training and Evaluation**

Three machine learning models were trained and evaluated using 5-fold cross-validation:

1. **Logistic Regression:**
   * Achieved a recall of 0.75, indicating moderate performance.
2. **Random Forest Classifier:**
   * Outperformed other models with a recall of 0.98, making it the best model for predicting employee turnover.
3. **Gradient Boosting Classifier:**
   * Also performed well with a recall of 0.93 but was slightly less effective than the Random Forest.

**Model Performance Metrics**

The ROC/AUC curves and confusion matrices were plotted for each model, confirming that the Random Forest Classifier outperformed the others.

**Retention Strategies**

Based on the predictions from the best model (Random Forest), employees were categorized into four risk zones, and retention strategies were suggested:

* **Safe Zone:** Continue current engagement practices.
* **Low Risk:** Monitor satisfaction levels and provide feedback.
* **Moderate Risk:** Implement targeted retention programs and career development opportunities.
* **High Risk:** Conduct stay interviews and address key concerns immediately.

**Conclusion**

The analysis provided valuable insights into the factors contributing to employee turnover at Portobello Tech. By leveraging machine learning models, the company can predict turnover and implement effective retention strategies to improve employee satisfaction and reduce turnover rates.