**Employee Turnover Analytics Report**

**Introduction**

Employee turnover is a significant challenge for organizations, reflecting the number of employees who leave a company over time. Portobello Tech, an innovative app developer, seeks to predict turnover by analyzing factors such as satisfaction levels, performance evaluations, work hours, project involvement, and promotions. This report outlines the data analysis and machine learning solutions designed to predict and manage employee turnover effectively.

**Objective**

The analysis aimed to achieve the following:

* Assess data quality and address potential deficiencies.
* Identify key factors contributing to employee turnover.
* Group employees into clusters based on satisfaction and evaluation levels.
* Handle class imbalance in the turnover dataset.
* Train and evaluate machine learning models for turnover prediction.
* Propose retention strategies tailored to specific employee groups.

**Data Quality Checks**

An initial review confirmed the dataset's integrity, with no missing values. Descriptive statistics provided a comprehensive view of numerical data distributions.

**Exploratory Data Analysis (EDA)**

* **Correlation Heatmap**: Revealed strong relationships among variables, particularly average monthly hours, satisfaction levels, and project counts.
* **Distribution Analysis**:
  + Employee satisfaction displayed a bimodal pattern, with peaks at both low and high levels.
  + Higher evaluation scores correlated with turnover.
  + Monthly hours were mostly within standard ranges, though outliers were common among employees who left.
* **Project Count Bar Plot**: Employees engaged in either very few or too many projects were more likely to leave, indicating an imbalance in workload.

**Clustering Analysis**

K-means clustering grouped employees who left into three categories based on satisfaction and evaluation levels:

1. Low satisfaction and low evaluation scores.
2. Moderate satisfaction but low evaluation scores.
3. High satisfaction and high evaluation scores, suggesting burnout.

**Class Imbalance Handling**

The dataset revealed a class imbalance, with fewer employees labeled as "left." To address this, SMOTE (Synthetic Minority Oversampling Technique) was applied, ensuring balanced training data and fair model performance.

**Machine Learning Model Training**

Three machine learning models were trained and assessed:

* **Logistic Regression**: Established a baseline with interpretable results.
* **Random Forest Classifier**: Showed robust performance across diverse data scenarios.
* **Gradient Boosting Classifier**: Excelled in identifying complex patterns within the data.

**Model Evaluation**

All models underwent 5-fold cross-validation and were evaluated on the test dataset using key metrics:

* **Classification Report**: Measured precision, recall, and F1-scores.
* **ROC/AUC**: Gradient Boosting achieved the highest AUC score, demonstrating superior predictive accuracy.
* **Confusion Matrix**: Offered detailed insights into prediction outcomes, balancing precision and recall.

**Retention Strategies**

Using the Gradient Boosting Classifier, employees were categorized into four risk zones based on their likelihood of leaving the company:

1. **Safe Zone (Green, <20%)**: Employees in this category have a very low risk of leaving. Current policies and practices are working well to retain them, and no immediate action is needed.
2. **Low-Risk Zone (Yellow, 20%-60%)**: These employees show some signs of dissatisfaction. Preventive measures, such as regular feedback sessions and improved engagement initiatives, can help address potential concerns early on.
3. **Medium-Risk Zone (Orange, 60%-90%)**: Employees in this group often deal with challenges like heavy workloads or limited career advancement opportunities. Retention efforts should focus on interventions such as mentorship programs, opportunities for professional development, or flexible work arrangements.
4. **High-Risk Zone (Red, >90%)**: Employees in this group are highly likely to leave. Immediate actions like one-on-one discussions, customized incentives, or addressing specific concerns are crucial to improving their retention.

**Conclusions and Recommendations**

* **Key Factors Influencing Turnover**:
  + Low levels of employee satisfaction.
  + Poor performance evaluations.
  + Workload imbalances, particularly for employees handling either too few or too many projects.
* **Best Model**:
* The Gradient Boosting Classifier was identified as the most accurate model for predicting turnover. It achieved the highest AUC score and outperformed other algorithms.
* **Retention Strategies**:
* Personalized retention plans for employees in the medium- and high-risk zones can significantly reduce turnover. By addressing their specific concerns and needs, the company can build a more engaged and satisfied workforce.

This analysis equips Portobello Tech’s HR team with actionable insights to better predict and manage employee turnover. By applying data-driven strategies, the organization can foster a healthier workplace, enhance employee satisfaction, and maintain a more committed and productive team.