**Employee Turnover Analytics: Predicting and Understanding Employee Churn**

**1. Introduction**

Employee turnover is a critical challenge for organizations, leading to significant costs associated with recruitment, training, and lost productivity. Understanding the factors that contribute to employees leaving and predicting who might leave can enable proactive interventions, thereby improving retention rates and fostering a more stable workforce. This project aims to analyze employee data to identify key drivers of turnover and build a predictive model to flag employees at risk of leaving, providing actionable insights for human resources and management.

**2. Exploratory Data Analysis (EDA) & Initial Insights**

Our initial exploration of the dataset revealed several crucial patterns:

* **Employee Satisfaction:** The distribution of satisfaction\_level was distinctly **bi-modal**, showing two large groups: one with very low satisfaction (around 0.1-0.15) and another with high satisfaction (0.7-0.9). This polarization suggests that employees are either quite unhappy or quite content, with fewer in the middle.
* **Last Evaluation Scores:** The last\_evaluation scores exhibited a **tri-modal distribution**, with peaks around 0.5-0.55, 0.8-0.85, and close to 1.0. This indicates distinct performance segments within the employee base.
* **Average Monthly Hours:** Similar to satisfaction, average\_monthly\_hours displayed a **bi-modal distribution**, with one group working standard hours (around 150-160) and another working significantly higher hours (around 250-270). This highlights potential workload imbalances.
* **Correlation with Turnover:** A **strong negative correlation (-0.39)** was observed between satisfaction\_level and left, indicating that lower satisfaction is a primary driver of turnover.
* **Projects vs. Turnover:** A critical insight emerged from the number\_project vs. Turnover analysis, revealing a **non-linear, U-shaped relationship**:
  + Employees with **2 projects** showed disproportionately high turnover (potential under-utilization/boredom).
  + Employees with **3-5 projects** had the lowest turnover (optimal workload).
  + Employees with **6-7 projects** experienced a sharply increasing turnover, with almost all employees with 7 projects leaving (strong indicator of burnout/overwork).

**3. Clustering Analysis of Employees Who Left**

To gain deeper insight into the profiles of employees who churned, KMeans clustering was applied to their satisfaction\_level and last\_evaluation scores, identifying three distinct groups:

* **Cluster 0 (Red - "Disengaged/Underperformers"):** Characterized by **low satisfaction (0.35-0.5)** and **mid-range evaluations (0.5-0.6)**. These employees might have left due to general unhappiness or feeling undervalued despite average performance.
* **Cluster 1 (Blue - "Overworked High-Performers/Burnout Victims"):** Exhibited **high satisfaction (0.7-1.0)** and **very high evaluations (0.8-1.0)**. This counter-intuitive group likely left due to burnout, unsustainable workload, or seeking a better work-life balance despite being top performers.
* **Cluster 2 (Green - "Truly Unhappy/Under-valued"):** Showed **extremely low satisfaction (0.0-0.2)** across a **wide range of evaluation scores (0.6-1.0)**. These individuals were deeply dissatisfied, regardless of their performance, pointing to potential issues with management or company culture.

This clustering highlights that turnover is not a singular problem but stems from diverse underlying issues, requiring tailored retention strategies.

**4. Methodology and Model Building**

To predict employee turnover (left variable), a supervised machine learning approach was adopted:

* **Data Preprocessing:** Categorical features (department, salary) were transformed using **one-hot encoding** (pd.get\_dummies) to convert them into a numerical format suitable for machine learning models.
* **Feature and Target Separation:** The dataset was split into features (X) and the target variable (y, representing left).
* **Addressing Class Imbalance:** The original dataset was imbalanced (7999 employees stayed, 2500 left). To prevent model bias towards the majority class, the **Synthetic Minority Over-sampling Technique (SMOTE)** was applied to the training data. This successfully balanced the classes to 7999 samples each.
* **Cross-Validation Strategy:** **Stratified K-Fold Cross-Validation (StratifiedKFold with n\_splits=5)** was employed. This ensures that each fold maintains the same proportion of employees who left as the original dataset, leading to a more robust and reliable evaluation of model performance, especially important given the initial class imbalance.
* **Models Evaluated:**
  + Logistic Regression
  + Random Forest Classifier
  + Gradient Boosting Classifier

**5. Model Evaluation and Selection**

The models were evaluated using a comprehensive set of metrics, including a Classification Report, ROC Curves, AUC scores, and Confusion Matrices.

* **Classification Report (Logistic Regression Example):** The initial Logistic Regression model (likely trained on SMOTE-resampled data) showed balanced performance across both classes, with precision, recall, and F1-scores ranging from 0.77 to 0.82, and an overall accuracy of 0.79. This indicated that the model was not biased towards the majority class.
* **ROC Curve Comparison:**
  + **Random Forest (AUC = 1.00):** Demonstrated perfect discriminative power, with its ROC curve hugging the top-left corner.
  + **Gradient Boosting (AUC = 0.99):** Also showed excellent discriminative power, very close to perfect.
  + **Logistic Regression (AUC = 0.86):** Performed well but was noticeably less effective at separating classes compared to the ensemble methods.

Based on the ROC curves and AUC scores, **Random Forest** was identified as the **best-performing model**, showcasing its superior ability to distinguish between employees who would leave and those who would stay. While an AUC of 1.00 is exceptional, it warrants cautious interpretation for potential overfitting or data leakage.

* **Confusion Matrix (for the Best Model - Random Forest):**
  + **True Negatives (TN): 2265** (Correctly predicted employees would stay)
  + **False Positives (FP): 21** (Incorrectly predicted employees would leave - false alarms)
  + **False Negatives (FN): 16** (Incorrectly predicted employees would stay when they left - missed opportunities)
  + **True Positives (TP): 698** (Correctly predicted employees would leave)

The confusion matrix confirms the model's high effectiveness. It correctly identified many both leavers and stayers, with very few false alarms (FP) or missed opportunities (FN). This indicates a highly reliable model for practical application.

**6. Risk Zone Implementation**

To make the model's predictions actionable, the continuous probabilities of leaving were categorized into four distinct "Risk Zones":

* **Safe Zone:** Employees with a very low probability of leaving.
  + **Recommendation:** Maintain current engagement strategies.
* **Low Risk Zone:** Employees with a slightly higher probability.
  + **Recommendation:** Monitor satisfaction levels and provide additional support.
* **Medium Risk Zone:** Employees with moderate probability.
  + **Recommendation:** Implement targeted retention strategies, such as career development opportunities.
* **High Risk Zone:** Employees with the highest probability of leaving.
  + **Recommendation:** Immediate intervention is required, such as one-on-one meetings to address concerns.

**Risk Zone Distribution:**

The distribution of employees across these zones was:

* **Safe Zone:** 1242 employees
* **High Risk Zone:** 646 employees
* **Low Risk Zone:** 111 employees
* **Medium Risk Zone:** 59 employees

This distribution highlights a polarization, with large groups in the "Safe" and "High Risk" categories, allowing for focused resource allocation.

**7. Conclusion & Recommendations**

This Employee Turnover Analytics project successfully built a robust predictive model, identified key drivers of churn, and translated complex predictions into actionable strategies. The Random Forest model demonstrated exceptional performance in predicting employee turnover.

**Key Recommendations for the Organization:**

1. **Prioritize High-Risk Employees:** Immediately engage with the **646 employees identified in the "High Risk" zone** through one-on-one meetings to understand and address their concerns.
2. **Optimize Workload:** Re-evaluate project assignments to ensure employees are within the **optimal 3-5 project range**. Address potential under-utilization for those with 2 projects and severe overwork for those with 6-7 projects.
3. **Address Satisfaction Extremes:** Investigate the root causes of **very low satisfaction** among employees (the "Disengaged" and "Truly Unhappy" clusters). This might involve surveying, exit interviews, or departmental reviews.
4. **Support High-Performers:** Implement strategies to prevent burnout among **high-performing employees (the "Overworked High-Performers" cluster)** who are leaving despite high satisfaction. This could include workload balancing, recognition programs, and promoting work-life balance.
5. **Proactive Monitoring:** Continue to monitor satisfaction levels for employees in the "Low Risk" zone and provide additional support to prevent them from escalating to higher risk categories.
6. **Maintain Engagement:** For the large "Safe Zone" group, continue existing positive engagement strategies that appear to be working effectively.