Write-up: Employee Turnover Prediction and Retention Strategies

1. Problem Statement

Employee turnover is a critical issue for organizations, leading to significant costs in recruitment, training, and lost productivity. The primary objective of this analysis was to understand the factors contributing to employee turnover, build a predictive model to identify employees at risk of leaving, and propose targeted retention strategies.

2. Data Loading and Quality Checks

- The dataset, HR_comma_sep.csv, was loaded into a Pandas DataFrame.
- Initial data quality checks revealed no missing values, indicating a clean dataset for analysis.

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand various aspects of the data and identify potential drivers of turnover:

- Employee Turnover Rate: Approximately 23.8% of employees left the company, indicating a significant turnover problem.
- Impact of Categorical Features:
    - Department ('sales'): Turnover rates varied across departments, with some showing higher proportions of employees leaving.
    - Salary: Employees with 'low' and 'medium' salaries had higher turnover rates compared to those with 'high' salaries.
    - Work Accident: Employees who had not experienced a work accident showed a slightly higher tendency to leave.
    - Promotion: Employees who had not been promoted in the last 5 years had a significantly higher turnover rate.
- Impact of Numerical Features:
    - Satisfaction Level: A clear bimodal distribution was observed. Employees who left tended to have either very low or moderately high satisfaction levels, suggesting complex underlying reasons.
    - Last Evaluation: Similar to satisfaction, employees who left showed peaks at very low and very high evaluation scores.
    - Average Monthly Hours: Employees working very low or very high average monthly hours were more prone to leaving, pointing to issues of underload or burnout.
    - Number of Projects: Both employees with too few (2) and too many (6-7) projects showed higher turnover, indicating optimal engagement with 3-5 projects.
    - Time Spend Company: Turnover was highest for employees with 3, 4, or 5 years of tenure.

- Clustering Employees Who Left: K-Means clustering (k=3) on satisfaction_level and last_evaluation for employees who left revealed three distinct groups:
    - Cluster 0 (Low Satisfaction, High Evaluation): Performing well but unhappy.
    - Cluster 1 (High Satisfaction, High Evaluation): Content but left, suggesting external factors (e.g., better opportunities).
    - Cluster 2 (Low Satisfaction, Low Evaluation): Dissatisfied and poor performance (expected turnover).

4. Data Transformation

- Categorical Encoding:
    - The 'sales' (department) column was one-hot encoded.
    - The 'salary' column was ordinal encoded ('low': 0, 'medium': 1, 'high': 2).

5. Handling Class Imbalance with SMOTE

- The target variable, 'left', was imbalanced (approximately 76% stayed, 24% left).
- The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to balance the classes, ensuring that the predictive models would not be biased towards the majority class.

6. Predictive Model Training and Evaluation

Three classification models were trained and evaluated using 5-fold stratified cross-validation on the balanced training data:

- Logistic Regression: Achieved F1-scores around 0.80 for both classes and an AUC of 0.87.
- Random Forest Classifier: Demonstrated very strong performance with F1-scores around 0.98 for both classes and an AUC of 0.99. The confusion matrix showed significantly fewer misclassifications.
- Gradient Boosting Classifier: Also performed strongly with F1-scores around 0.96 for both classes and an AUC of 0.99. The confusion matrix indicated excellent performance.

7. Model Comparison and Best Model Selection

- ROC/AUC Curves and Confusion Matrices: Both Random Forest and Gradient Boosting models showed superior discriminative power and significantly fewer misclassifications compared to Logistic Regression.
- Best Model: The Random Forest Classifier was selected as the best model due to its consistently high F1-scores (0.98) and AUC (0.99), marginally outperforming Gradient Boosting and significantly surpassing Logistic Regression. While Gradient Boosting was close, Random Forest offered a slightly better balance in performance metrics.

8. Recall vs. Precision for Turnover Prediction

- Recall was identified as the more critical metric for employee turnover prediction. The cost of a False Negative (failing to identify an employee who will leave) is generally much higher (e.g., loss of talent, knowledge, productivity) than the cost of a False Positive (intervening with an employee who would have stayed). Prioritizing recall ensures that most at-risk employees are identified for proactive intervention.

9. Predicting Turnover Probabilities and Categorizing Employees

- The best model (Random Forest Classifier) was used to predict the probability of turnover for each employee in the test set.
- Employees were then categorized into four risk zones based on these probabilities:
    - Safe Zone (Green): Predicted Turnover Probability < 20%
    - Low-Risk Zone (Yellow): 20% <= Probability < 60%
    - Medium-Risk Zone (Orange): 60% <= Probability < 80%
    - High-Risk Zone (Red): Predicted Turnover Probability >= 80%

10. Retention Strategies by Turnover Zone

Targeted retention strategies were proposed for each zone:

- Safe Zone (Green): Maintain engagement, recognize achievements, provide growth opportunities, and leverage them as mentors.
- Low-Risk Zone (Yellow): Proactive engagement, "stay interviews," mentorship, career development, and competitive compensation reviews.
- Medium-Risk Zone (Orange): Aggressive retention efforts, direct manager intervention, personalized career coaching, workload review, and addressing specific grievances.
- High-Risk Zone (Red): Focus on knowledge transfer if departure is inevitable; for undecided cases, consider significant incentives or changes, and conduct thorough exit interviews.

Conclusion

This analysis provides a data-driven framework for understanding and addressing employee turnover. By leveraging predictive modeling and tailored retention strategies, organizations can significantly improve their ability to retain valuable talent and foster a more stable and engaged workforce.