Employee Turnover Analytics

Project Summary

This analysis examined employee turnover patterns using machine learning techniques on HR data containing 14,999 employee records. The project aimed to identify key factors contributing to employee departures and develop predictive models for retention strategies.

Key Findings

Data Quality & Characteristics

- **No missing values** in the dataset
- **5,346 duplicate rows** identified (35.6% of data)
- Statistical testing confirmed duplicates represent distinct employee subgroups, not data errors
- Class imbalance: 76.2% stayed vs 23.8% left

Turnover Factors

Work-Life Balance is Critical:

- Employees with 2 projects (underutilized) and 6-7 projects (overworked) show highest turnover
- Optimal project count: 3-5 projects
- Average monthly hours: 201 hours (20-25% above standard 160-176 hours)

Satisfaction Patterns:

- Company average satisfaction: 0.61/1.0
- Salary has minimal impact on satisfaction (High: 0.64, Medium: 0.62, Low: 0.60)
- Overtime doesn't correlate with salary levels

Employee Segmentation (K-means Clustering)

Three distinct groups among employees who left:

1. **Underperformers** (41% satisfaction, 52% evaluation) - Need development support
2. **High Performers** (81% satisfaction, 91% evaluation) - Successful retention examples
3. **Burned Out Stars** (11% satisfaction, 87% evaluation) - High performers leaving due to dissatisfaction

Model Performance

Best Model: Gradient Boosting Classifier

- **Accuracy: 97%**
- **Precision: 97.3%** (few false positives)

- **Recall: 93.3%** (catches most actual departures)
- **AUC: 0.989** (near-perfect discrimination)

Model Comparison

Model	Accuracy	Precision	Recall	AUC
Logistic Regression	84%	70.6%	53.2%	0.821
Random Forest	90%	97.4%	47.9%	0.953
Gradient Boosting	**97%**	**97.3%**	**93.3%**	**0.989**

Risk-Based Retention Strategy

Employee Risk Zones

- **Safe Zone (Green)**: <20% risk - Continue current support
- **Low-Risk (Yellow)**: 20-60% risk - Engagement initiatives
- **Medium-Risk (Orange)**: 60-90% risk - Manager interventions
- **High-Risk (Red)**: >90% risk - Urgent HR action

Actionable Recommendations

Immediate Actions:

1. **Workload Management**: Ensure 3-5 projects per employee
2. **Identify Burned-Out Stars**: Focus on high-performing but dissatisfied employees
3. **Implement Risk Monitoring**: Use predictive model for early intervention

Strategic Initiatives:

1. **Work-Life Balance Programs**: Address 20-25% overtime issue
2. **Career Development**: Prevent underutilization of 2-project employees
3. **Recognition Systems**: Retain high performers before they become dissatisfied

Business Impact

The predictive model enables proactive retention management, potentially reducing turnover costs and retaining valuable talent. The 97% accuracy provides reliable early warning for HR interventions, while the risk-based approach allows for targeted resource allocation.

Technical Methodology

- **Data Preprocessing**: Handled duplicates, class imbalance with SMOTE
- **Feature Engineering**: Categorical encoding, scaling
- **Clustering:** K-means Clustering
- **Cross-Validation**: 5-fold stratified approach with duplicate-aware grouping
- **Evaluation**: Multiple metrics prioritizing recall for early intervention