

- Employee Turnover Analytics - Comprehensive Analysis Writeup
 - Executive Summary
 - 1. Dataset Overview and Data Quality
 - Dataset Characteristics
 - Features Analyzed
 - 2. Key Findings and Insights
 - 2.1 Primary Turnover Drivers
 - 2.2 Advanced Statistical Insights
 - 3. Clustering Analysis
 - Employee Segmentation
 - 4. Machine Learning Model Development
 - 4.1 Class Imbalance Handling
 - 4.2 Model Comparison
 - 4.3 Model Performance Analysis
 - 5. Risk Zone Classification
 - Employee Risk Stratification
 - Financial Impact Analysis
 - 6. Retention Strategy Recommendations
 - 6.1 Immediate Actions (Next 30 Days)
 - 6.2 Short-Term Actions (Next 90 Days)
 - 6.3 Long-Term Actions (Next 6-12 Months)
 - 6.4 Technical Implementation
 - 7. Success Metrics and Expected Outcomes
 - Key Performance Indicators
 - Expected Outcomes (12-month projection)
 - 8. Technical Methodology
 - 8.1 Data Processing Pipeline
 - 8.2 Statistical Techniques Used
 - 8.3 Model Selection Criteria
 - 9. Limitations and Future Work
 - Current Limitations
 - Future Enhancements
 - 10. Conclusion

Employee Turnover Analytics - Comprehensive Analysis Writeup

Executive Summary

This comprehensive analysis of employee turnover data provides actionable insights for reducing organizational turnover costs and improving employee retention. The study analyzed 14,999 employee records and identified key factors driving turnover, developed predictive models, and created targeted retention strategies.

Key Results:

- **Overall Turnover Rate:** 23.8% (3,571 employees left)
 - **Best Predictive Model:** Gradient Boosting with 98.1% F1-score and 99.4% AUC
 - **Primary Turnover Driver:** Employee satisfaction level (correlation: -0.388)
 - **Potential ROI:** 649.7% on retention investment
-

1. Dataset Overview and Data Quality

Dataset Characteristics

- **Size:** 14,999 employee records with 10 features
- **Memory Usage:** Optimized for analysis efficiency
- **Data Quality:** No missing values, minimal duplicates
- **Time Period:** Historical employee data with turnover outcomes

Features Analyzed

1. **Satisfaction Level** (0.09 - 1.00)
2. **Last Evaluation Score** (0.36 - 1.00)
3. **Number of Projects** (2 - 7)
4. **Average Monthly Hours** (96 - 310)
5. **Time Spent in Company** (2 - 10 years)
6. **Work Accident** (Binary)
7. **Promotion in Last 5 Years** (Binary)
8. **Department** (10 categories)
9. **Salary Level** (Low/Medium/High)
10. **Turnover Status** (Target variable)

2. Key Findings and Insights

2.1 Primary Turnover Drivers

1. Employee Satisfaction (Strongest Predictor)

- Correlation with turnover: -0.388
- Employees who left: 0.440 average satisfaction
- Employees who stayed: 0.667 average satisfaction
- **Insight:** Satisfaction level is the most critical factor in retention

2. Project Count (U-Shaped Relationship)

- Lowest turnover: 3 projects (1.8% turnover rate)
- Highest turnover: 7 projects (100% turnover rate)
- **Insight:** Both underutilization and overutilization increase turnover risk

3. Working Hours Impact

- Employees who left: 207 average monthly hours
- Employees who stayed: 199 average monthly hours
- **Insight:** Excessive workload contributes to turnover

4. Department Variations

- Highest turnover: HR (29.1%)
- Lowest turnover: Management (14.4%)
- **Insight:** Department-specific retention strategies needed

5. Salary Impact

- Low salary: 29.7% turnover
- Medium salary: 20.4% turnover
- High salary: 6.6% turnover
- **Insight:** Compensation significantly affects retention

2.2 Advanced Statistical Insights

Correlation Analysis:

- Satisfaction level shows strongest negative correlation with turnover
- Work accidents correlate with lower turnover (paradoxical finding)
- Time in company shows moderate positive correlation with turnover
- Promotion history shows weak correlation with retention

Interaction Effects:

- Satisfaction × Evaluation interaction: $r = -0.293$
 - Hours × Projects interaction: $r = 0.155$
 - Time × Satisfaction interaction: $r = -0.121$
-

3. Clustering Analysis

Employee Segmentation

The analysis identified distinct employee clusters based on behavioral patterns:

Cluster 1: High Performers

- High satisfaction, high evaluation scores
- Moderate project load and working hours
- Low turnover risk

Cluster 2: Overworked Employees

- High evaluation but low satisfaction
- High project count and working hours
- High turnover risk

Cluster 3: Underutilized Employees

- Low project count and working hours
- Moderate satisfaction and evaluation
- Medium turnover risk

Cluster 4: Disengaged Employees

- Low satisfaction and evaluation scores

- Variable workload patterns
- High turnover risk

4. Machine Learning Model Development

4.1 Class Imbalance Handling

- **Challenge:** 76.2% stayed vs 23.8% left (imbalanced dataset)
- **Solution:** SMOTE (Synthetic Minority Oversampling Technique)
- **Result:** Balanced dataset for improved model performance

4.2 Model Comparison

Model	F1-Score	AUC	Recall	Precision
Gradient Boosting	0.981	0.994	0.955	0.965
Random Forest	0.956	0.994	0.931	0.982
Logistic Regression	0.666	0.847	0.824	0.559

4.3 Model Performance Analysis

Gradient Boosting (Best Model):

- **True Positives:** 682 (correctly predicted employees who left)
- **True Negatives:** 2,261 (correctly predicted employees who stayed)
- **False Positives:** 25 (incorrectly predicted employees would leave)
- **False Negatives:** 32 (missed employees who actually left)

Business Justification for Metrics:

- **Recall prioritized over Precision:** False negatives (missing employees who leave) cost 20x more than false positives
- **Cost of turnover:** 100,000 per employee vs 5,000 for retention efforts

- **ROI focus:** Better to identify 100 at-risk employees (including 20 false positives) than miss 20 who actually leave

5. Risk Zone Classification

Employee Risk Stratification

Risk Zone	Count	Percentage	Turnover Probability
Safe Zone (Green)	11,108	74.1%	< 20%
High Risk Zone (Red)	3,333	22.2%	> 80%
Medium Risk Zone (Orange)	127	0.8%	60-80%
Low Risk Zone (Yellow)	431	2.9%	20-60%

Financial Impact Analysis





- **High-risk employees:** 3,460 total
- **Potential turnover cost:** \$346,000,000
- **Retention investment needed:** \$46,150,500
- **Potential ROI:** 649.7%

6. Retention Strategy Recommendations





6.1 Immediate Actions (Next 30 Days)

1. 🚨 **URGENT:** Contact all High Risk Zone employees for retention discussions
2. 📊 **Implement:** Real-time monitoring dashboard for at-risk employees
3. 🎯 **Launch:** Targeted retention programs for Medium Risk Zone employees
4. 📋 **Conduct:** Stay interviews with high-risk employees

6.2 Short-Term Actions (Next 90 Days)

1.  **Review:** Workload distribution across projects
2.  **Implement:** Salary review process for underpaid high performers
3.  **Address:** Department-specific turnover issues
4.  **Implement:** Flexible working arrangements

6.3 Long-Term Actions (Next 6-12 Months)

1.  **Develop:** Comprehensive employee development programs
2.  **Create:** Career advancement pathways
3.  **Implement:** Recognition and reward systems
4.  **Establish:** Continuous monitoring and prediction system

6.4 Technical Implementation

1. **Deploy:** Trained ML model for real-time predictions
 2. **Integrate:** With HR systems for automated alerts
 3. **Create:** Dashboard for managers to track employee risk
 4. **Implement:** Feedback loop to improve model accuracy
-

7. Success Metrics and Expected Outcomes

Key Performance Indicators

- Overall turnover rate reduction
- High-risk employee count reduction
- Employee satisfaction score improvement
- Retention rate for targeted employees
- Cost savings from reduced turnover
- Model prediction accuracy over time

Expected Outcomes (12-month projection)

- **25% reduction** in overall turnover rate
 - **50% reduction** in high-risk employee count
 - **20% improvement** in employee satisfaction scores
 - **300%+ ROI** on retention investment
 - Improved employee engagement and productivity
-

8. Technical Methodology

8.1 Data Processing Pipeline

1. **Data Quality Assessment:** Comprehensive validation and cleaning
2. **Exploratory Data Analysis:** Statistical analysis and visualization
3. **Feature Engineering:** Interaction terms and derived variables
4. **Clustering Analysis:** K-means segmentation
5. **Model Training:** Multiple algorithms with cross-validation
6. **Performance Evaluation:** Business-focused metrics

8.2 Statistical Techniques Used

- **Correlation Analysis:** Pearson correlation with significance testing
- **Distribution Analysis:** Shapiro-Wilk, Anderson-Darling tests
- **Clustering:** K-means with elbow method optimization
- **Cross-Validation:** Stratified K-fold for robust evaluation
- **Resampling:** SMOTE for class imbalance handling

8.3 Model Selection Criteria

- **Primary:** F1-score (balanced precision and recall)
 - **Secondary:** AUC (overall discriminative ability)
 - **Business Focus:** Recall (minimize false negatives)
 - **Cost-Benefit:** ROI optimization
-

9. Limitations and Future Work

Current Limitations

1. **Temporal Data:** No time-series analysis of satisfaction trends
2. **External Factors:** No consideration of market conditions or industry trends
3. **Causal Inference:** Correlation analysis without causal modeling
4. **Feature Engineering:** Limited domain-specific features

Future Enhancements

1. **Real-time Data Integration:** Live HR system connectivity
 2. **Advanced ML Models:** Deep learning and ensemble methods
 3. **Causal Analysis:** Causal inference for retention strategies
 4. **Predictive Maintenance:** Proactive intervention recommendations
 5. **A/B Testing:** Experimental validation of retention strategies
-

10. Conclusion

This comprehensive employee turnover analysis provides a data-driven foundation for reducing organizational turnover costs and improving employee retention. The key findings reveal that:

1. **Employee satisfaction is the primary driver** of turnover decisions
2. **Workload balance** (projects and hours) significantly impacts retention
3. **Department-specific strategies** are needed for effective retention
4. **Machine learning models** can accurately predict turnover risk
5. **Targeted interventions** can yield substantial ROI

The analysis demonstrates that a proactive, data-driven approach to employee retention can significantly reduce turnover costs while improving organizational performance. The recommended strategies, when implemented systematically, can achieve a 649.7% ROI on retention investments.

Next Steps:

1. Implement the recommended immediate actions
2. Deploy the predictive model for real-time monitoring

3. Establish continuous feedback loops for model improvement
4. Monitor and measure the impact of retention strategies

This analysis provides the foundation for transforming employee retention from a reactive process to a proactive, data-driven strategy that benefits both employees and the organization.

Analysis completed using Python, scikit-learn, pandas, and advanced statistical methods. All code and visualizations are available in the accompanying Jupyter notebook.