Employee Attrition Analysis

Milestone 2: Advanced Data Analysis and Feature Engineering

Executive Summary

This report details the advanced data analysis and feature engineering phase of the Employee Attrition Analysis project. Building on the cleaned dataset from Milestone 1, we conducted comprehensive statistical testing, correlation analysis, and feature engineering to identify the most significant predictors of employee attrition.

Chi-square tests revealed that OverTime, JobRole, JobLevel, and MaritalStatus are strongly associated with attrition, while Education, Gender, PerformanceRating, and RelationshipSatisfaction showed no significant relationship.

Key Findings:

- Chi-square tests identified 11 features significantly associated with attrition ($p < 0.05$)
- OverTime showed the strongest association ($\chi^2 = 87.56$, $p < 0.001$)
- Four features deemed independent of attrition: Education, Gender, PerformanceRating, RelationshipSatisfaction
- Feature selection reduced predictor set from 30 to 26 meaningful features
- Statistical analysis confirmed patterns observed in EDA

1. Advanced Statistical Analysis

1.1 Chi-Square Test of Independence

To assess the relationship between categorical features and employee attrition, we conducted chi-square tests of independence. This statistical test determines whether there is a significant association
between two categorical variables.

Methodology:

- Null Hypothesis (H_o): Feature and Attrition are independent
- Alternative Hypothesis (H_1): Feature and Attrition are associated
- Significance Level: $\alpha = 0.05$
- Test Applied: Chi-square test of independence using scipy.stats

1.2 Chi-Square Test Results

The following table presents chi-square test statistics and p-values for all categorical features tested against the Attrition variable, sorted by p-value (most significant first):

Rank	Feature	Chi-Square Statistic	P-Value	Significance
1	OverTime	87.56	<0.001	***
2	JobRole	86.19	<0.001	***
3	JobLevel	72.53	<0.001	***
4	MaritalStatus	46.16	<0.001	***
5	JobInvolvement	28.49	<0.001	***
6	BusinessTravel	24.18	<0.001	***
7	EnvironmentSatisfaction	22.50	<0.001	***
8	JobSatisfaction	17.51	<0.001	***
9	WorkLifeBalance	16.33	<0.001	***
10	Department	10.80	0.005	**
11	EducationField	16.02	0.007	**
12	RelationshipSatisfaction	5.24	0.155	NS
13	Gender	1.12	0.291	NS
14	Education	3.07	0.545	NS
15	PerformanceRating	0.00	0.990	NS

Legend: *** $p < 0.001$ (highly significant), ** $p < 0.01$ (significant), NS = Not Significant ($p \geq 0.05$)

1.3 Statistical Findings Interpretation

Highly Significant Features ($p < 0.001$):

- **OverTime ($\chi^2 = 87.56$):** Strongest predictor. Employees working overtime have significantly different attrition rates.
- **JobRole ($\chi^2 = 86.19$):** Job role strongly influences attrition. Certain roles show markedly higher turnover.
- **JobLevel ($\chi^2 = 72.53$):** Career level is crucial. Entry and junior levels show elevated attrition.
- **MaritalStatus ($\chi^2 = 46.16$):** Marital status impacts retention. Single employees more likely to leave.
- **JobInvolvement ($\chi^2 = 28.49$):** Employee engagement level directly affects retention decisions.
- **BusinessTravel ($\chi^2 = 24.18$):** Travel frequency impacts work-life balance and attrition.
- **EnvironmentSatisfaction ($\chi^2 = 22.50$):** Workplace environment quality influences employee retention.
- **JobSatisfaction ($\chi^2 = 17.51$):** Job satisfaction is a key driver of staying vs. leaving.
- **WorkLifeBalance ($\chi^2 = 16.33$):** Work-life balance significantly impacts attrition decisions.

Moderately Significant Features ($0.001 < p < 0.05$):

- **Department ($\chi^2 = 10.80$, $p = 0.005$):** Department affiliation shows significant but weaker association.
- **EducationField ($\chi^2 = 16.02$, $p = 0.007$):** Field of study demonstrates meaningful relationship with attrition.

Non-Significant Features ($p \geq 0.05$):

The following features showed NO significant association with attrition and are candidates for removal
from the predictive model:

- **RelationshipSatisfaction ($p = 0.155$):** No evidence of relationship with attrition
- **Gender ($p = 0.291$):** Gender-neutral attrition patterns
- **Education ($p = 0.545$):** Education level does not significantly impact retention
- **PerformanceRating ($p = 0.990$):** No association (nearly all rated "Excellent")

2. Feature Selection Strategy

2.1 Statistical Feature Selection

Based on chi-square test results, we implemented a systematic feature selection strategy to optimize the predictive model by removing features with no significant association with attrition ($p \geq 0.05$).

Features Recommended for Removal:

Feature	P-Value	Reason for Removal	Impact
Education	0.545	No significant association with attrition	Reduces noise
Gender	0.291	Gender-neutral attrition patterns	Promotes fairness
PerformanceRating	0.990	Lack of variance (most rated "Excellent")	No predictive value
RelationshipSatisfaction	0.155	Borderline insignificant, weak relationship	Marginal benefit

Result: Feature set reduced from 30 predictors to 26 statistically significant features, improving model efficiency and interpretability.

2.2 Retained Features Summary

The final feature set includes 26 predictors across multiple categories:

Categorical Features (11): OverTime, JobRole, JobLevel, MaritalStatus, JobInvolvement, BusinessTravel, EnvironmentSatisfaction, JobSatisfaction, WorkLifeBalance, Department, EducationField

Numerical Features (15): Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager

3. Feature Engineering

3.1 Label Encoding Implementation

During Milestone 1, we successfully transformed 8 ordinal features from numeric codes to meaningful categorical labels. This transformation enhances:

- Clearer visualization and reporting
- Improved data quality documentation

4. Advanced Data Visualization Insights

4.1 Attrition Patterns by Key Features

Advanced visualization analysis revealed distinct patterns in employee attrition across different categorical and numerical features:

OverTime and Attrition:

- Employees working overtime: ~30% attrition rate
- Employees not working overtime: ~10% attrition rate
- 3x higher attrition for overtime workers
- Strongest single predictor identified

Job Level Distribution:

- Entry Level: Highest attrition (~26%)
- Junior Level: Moderate-high attrition (~20%)
- Mid Level: Moderate attrition (~15%)
- Senior & Executive: Lowest attrition (<5%)
- Clear inverse relationship between level and attrition

4.2 Numerical Feature Distributions

Analysis of numerical features revealed:

- Age: Younger employees (18-30) show higher attrition
- Monthly Income: Strong negative correlation with attrition
- Distance from Home: Slight positive correlation with leaving
- Years at Company: Employees with 0-2 years tenure most likely to leave
- Total Working Years: More experienced workers tend to stay

5. Correlation Analysis

5.1 Numerical Feature Correlations

Correlation analysis among numerical features identified several important relationships:

Strong Positive Correlations:

- JobLevel ↔ MonthlyIncome ($r \approx 0.95$): Expected relationship
- TotalWorkingYears ↔ Age ($r \approx 0.68$): Natural life-stage correlation
- YearsAtCompany ↔ YearsInCurrentRole ($r \approx 0.76$): Tenure consistency
- YearsAtCompany ↔ YearsWithCurrManager ($r \approx 0.77$): Management stability

Multicollinearity Concerns:

High correlations between predictors may cause multicollinearity issues in modeling.
Consider the following strategies:

- Use tree-based models (Random Forest, XGBoost) which handle multicollinearity well

6. Key Insights and Actionable Findings

6.1 Critical Attrition Drivers

Driver	Recommended Actions
OverTime (Highest Impact)	Reduce mandatory overtime, implement rotation policies, provide overtime compensation
Job Level	Focus retention efforts on entry/junior employees, create clear career paths, mentorship programs
Marital Status	Develop family-friendly policies, consider life-stage benefits, flexible scheduling
Work-Life Balance	Promote healthy work culture, enforce time-off policies, remote work options
Job Satisfaction	Regular feedback sessions, address workplace concerns, recognition programs

7. Conclusions and Next Steps

7.1 Summary of Achievements

Milestone 2 successfully delivered comprehensive statistical analysis and feature engineering insights. Through rigorous chi-square testing, we identified the most significant predictors of employee attrition and eliminated non-contributing features. The analysis revealed that overtime work, job characteristics, and work-life factors are the primary drivers of attrition, while demographic factors like gender and education level show no significant impact.

- Conducted chi-square tests on 15 categorical features
- Identified 11 statistically significant predictors
- Removed 4 non-significant features (Education, Gender, PerformanceRating, RelationshipSatisfaction)

7.2 Readiness for Predictive Modeling

The dataset is now fully prepared for predictive modeling with:

- ✓ Clean, validated data (1,470 records)
- ✓ Optimized feature set (26 significant predictors)
- ✓ Statistical validation of feature importance
- ✓ Understanding of class imbalance (16% attrition)

7.3 Recommended Next Steps

Phase 3: Model Development

- Address class imbalance using SMOTE or class weights
- Create derived features in modeling pipeline
- Train multiple algorithms (Logistic Regression, Random Forest, XGBoost)
- Evaluate models using appropriate metrics (AUC-ROC, Precision, Recall, F1)

Phase 4: Model Deployment

- Select best-performing model
- Create prediction pipeline
- Build interactive dashboard
- Implement monitoring and retraining strategy

Conclusion: The advanced data analysis and feature engineering phase has established a solid foundation for building accurate predictive models. With statistically validated features and deep understanding of attrition patterns, the project is well-positioned to deliver actionable insights that will help the organization proactively manage employee retention.