Employee Attrition Analysis

Which workplace factors most strongly influence employee attrition, and how can organizations leverage these factors to reduce turnover?

1. Executive Summary

Employee attrition represents a significant operational and financial challenge for large organizations, particularly when turnover is driven by preventable workplace factors rather than unavoidable personal circumstances. Using IBM’s HR analytics dataset, this project examines which workplace characteristics most strongly influence an employee’s likelihood of leaving and how organizations can leverage these insights to reduce turnover risk.

Our analysis finds that workload pressure and compensation are the dominant drivers of attrition. Employees who consistently work overtime are substantially more likely to leave, while higher monthly income and longer tenure significantly reduce attrition risk. Engagement-related factors such as job involvement, job satisfaction, and work–life balance also play an important protective role. These findings are consistent across both logistic regression and decision tree models, strengthening confidence in their robustness.

2. Introduction and Motivation

Employee attrition is a persistent challenge for large organizations, leading to increased recruitment costs, productivity loss, and disruption to team continuity. Industry estimates suggest that replacing an employee can cost between 50% and 200% of their annual salary, making turnover reduction a meaningful source of organizational value.

This project focuses on IBM, a large and diversified organization whose publicly available HR dataset captures employee characteristics related to compensation, workload, engagement, and career progression. The objective is to provide data-driven insights into which factors most strongly influence attrition and how organizations can realistically intervene.

3. Data Wrangling Process

3.1 Data Identification and Ingestion

The dataset used in this project was obtained from Kaggle's *IBM HR Analytics Employee Attrition* dataset and downloaded as a CSV file (see Exhibit 1). It contains over 35 employee-level variables covering demographics, job characteristics, compensation, performance, and satisfaction. Data ingestion was performed using Python's pandas library, with no web scraping, API access, or SQL queries required.

3.2 Variable Selection and Data Reduction

The data wrangling process focused on isolating measurable and actionable workplace factors while improving interpretability. Variables were retained based on relevance to attrition theory, managerial actionability, and statistical usefulness. The final set included Age, MonthlyIncome, OverTime, JobSatisfaction, WorkLifeBalance, JobInvolvement, YearsAtCompany, and YearsSinceLastPromotion. Highly correlated tenure measures were removed in favor of YearsAtCompany, while demographic, job-specific, low-variance, and redundant variables were excluded to simplify the analysis and improve generalizability.

3.3 Data Cleaning, Encoding, and Descriptive Analytics

Categorical variables were converted into numerical form, with binary variables encoded as indicators and ordinal variables preserved to retain rank information. The dataset contained no missing values for the selected variables, so no imputation was required. All variables were reviewed for consistency and appropriate data types prior to modeling (see Exhibit 1).

Descriptive analytics examined variable distributions and their relationships with attrition. Univariate analyses included distributions of age, income, overtime, and attrition status, while bivariate analysis focused on correlation patterns among workplace factors (see Exhibit 2). These summaries informed subsequent modeling decisions.

4. Data Analytics and Insights

4.1 Model Selection

Two complementary models were used to analyze employee attrition. Logistic regression quantifies the marginal effect of each workplace factor on attrition probability, while a decision tree captures non-linear relationships and provides an interpretable ranking of feature importance. Supporting visuals include a correlation heatmap and a decision tree feature importance chart highlighting key attrition drivers.

4.2 Logistic Regression Results

The reduced logistic regression model achieves an accuracy of 85.3%. Although influenced by class imbalance, the model shows high specificity (98.3%), indicating strong performance in identifying employees who remain. Sensitivity is lower at 17.7%, but when attrition is predicted, the model is correct approximately 67% of the time, making it well suited for targeted retention efforts.

Overtime exhibits the strongest positive correlation with attrition, while monthly income, years at company, and job involvement are negatively correlated (Exhitbit 2). At the 5% significance level, OverTime is the most influential predictor, with employees working overtime being over four times as likely to leave. Higher income, stronger job involvement, better job satisfaction, improved work–life balance, and longer tenure significantly reduce attrition risk, while more years since last promotion slightly increase it. Exhibit 3 provides detailed coefficient estimates and odds ratios.

4.3 Decision Tree Insights

The decision tree model achieves a test accuracy of 83.9%, comparable to the logistic regression model. Feature importance analysis confirms the regression results and adds interpretability. As shown in Exhibit 4, monthly income, age, overtime status, and years at company are the most influential predictors of attrition in the non-linear model. These results reinforce the conclusion that compensation and workload pressure dominate attrition risk, while engagement-related factors provide meaningful secondary protection.

4.4 Hypothesis Evaluation

The findings are consistent across both models, supporting all core hypotheses. Overtime significantly increases attrition risk, with an odds ratio of 4.39. Higher compensation reduces attrition, while engagement factors such as job involvement, job satisfaction, and work–life balance consistently mitigate turnover risk.

5. Discussion and Managerial Implications

This analysis shows that employee attrition is driven by measurable workplace factors rather than random individual decisions. Across both logistic regression and decision tree models, overtime, compensation, tenure, promotion timing, and engagement consistently influence attrition risk.

Overtime is the strongest predictor of attrition, with employees working overtime over four times more likely to leave. Higher compensation substantially reduces attrition risk and ranks as the most influential feature in the decision tree. Longer tenure lowers attrition probability, while extended time since last promotion increases it. Engagement-related factors, including job involvement and work–life balance, further mitigate attrition risk.

These findings suggest clear managerial priorities. Organizations should focus first on reducing sustained overtime, particularly among employees with high workload exposure, followed by targeted compensation reviews and clearer promotion pathways. Engagement initiatives can provide additional, lower-cost retention benefits.

Reducing overtime exposure among the top 20% of affected employees could plausibly lower overall attrition by approximately 3–5%. Given that replacing an employee typically costs 50%–200%

of annual salary, even modest reductions in turnover could generate meaningful cost savings for IBM, with long-term benefits likely outweighing short-term trade-offs.

Exhibits

Exhibit 1: Code for variable selection, missing value check, and categorical variable encoding

```
# Load and inspect raw dataset
df = pd.read_csv("employee_attrition_data.csv")
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Age                      1470 non-null   int64
 1   MonthlyIncome            1470 non-null   int64
 2   OverTime                 1470 non-null   object
 3   JobSatisfaction          1470 non-null   int64
 4   WorkLifeBalance          1470 non-null   int64
 5   JobInvolvement           1470 non-null   int64
 6   YearsAtCompany           1470 non-null   int64
 7   YearsSinceLastPromotion  1470 non-null   int64
 8   Attrition                1470 non-null   object


# Encode categorical variables and create final dataset
df_clean = pd.get_dummies(
   df,
   columns=["OverTime", "Attrition"],
   drop_first=True,
   dtype=int
)

df_clean.head()
```

	Age	**MonthlyIncome**	**JobSatisfaction**	**WorkLifeBalance**	**JobInvolvement**	**YearsAtCompany**	**YearsSinceLastPromotion**	**OverTime_Yes**	**Attrition_Yes**
0	41	5993	4	1	3	6	0	1	1
1	49	5130	2	3	2	10	1	0	0
2	37	2090	3	3	2	0	0	1	1

3	33	2909	3	3	3	8	3	1	0
4	27	3468	2	3	3	2	2	0	0

Exhibit 2: Heatmap of pairwise correlations among selected employee characteristics

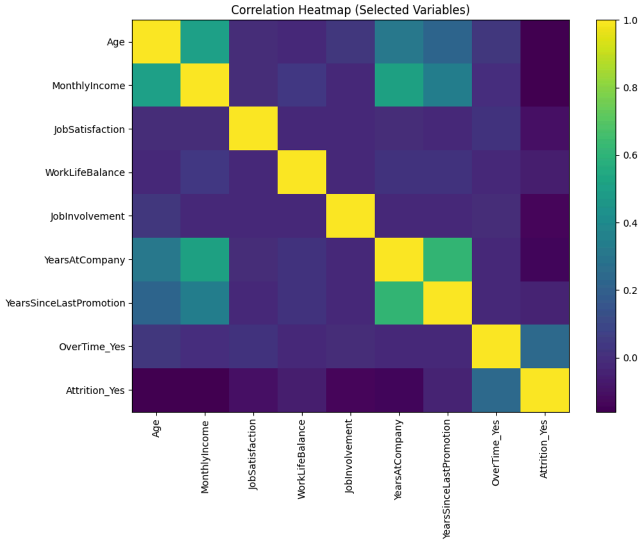


Exhibit 3: Logistic Regression Results

Panel A: Coefficient Estimates and Odds Ratios

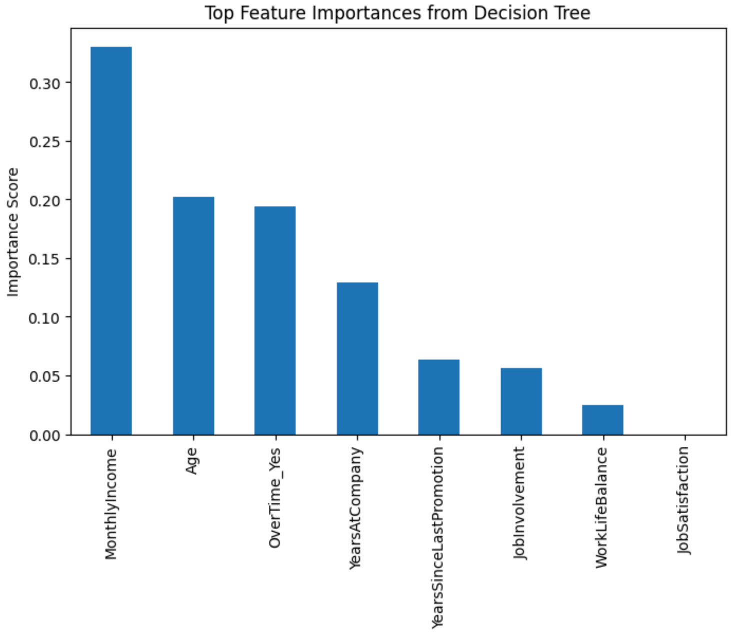
Coefficient	Odds_Ratio	p_value	
JobSatisfaction	**-0.318**	**0.728**	**0.0000**
JobInvolvement	**-0.527**	**0.591**	**0.0000**
OverTime_Yes	**1.479**	**4.389**	**0.0000**
Age	**-0.035**	**0.966**	**0.0006**
YearsAtCompany	**-0.082**	**0.921**	**0.0006**
YearsSinceLastPromotion	**0.118**	**1.125**	**0.0010**
MonthlyIncome	**-0.000**	**1.000**	**0.0024**
WorkLifeBalance	**-0.268**	**0.765**	**0.0125**

Panel B: Model Performance

```
Confusion Matrix (rows=true, cols=pred):
[[363   7]
 [ 57  14]]
```

	Metric	Value
0	Accuracy	0.855
1	Sensitivity (Attrition=1)	0.197
2	Specificity (Attrition=0)	0.981

Exhibit 4: Feature importance scores



Work Cited

Subhash, P. (2017). IBM HR Analytics Employee Attrition & Performance. Www.kaggle.com. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Yi, R. (2024, September 18). Employee Retention Depends on Getting Recognition Right. Gallup; Gallup.https://www.gallup.com/workplace/650174/employee-retention-depends-getting-recognition-right.aspx