**DAT‑430 Project Two Predicting HR Attrition**  
**Rand M Codon II**

**Executive Summary**

In this analysis, I leveraged HR attrition data to quantify the drivers of employee turnover and build predictive models to estimate who is most likely to leave. I established a baseline attrition rate of **29.60%,** conducted exploratory data analysis (EDA) to surface key patterns, engineered features for modeling, and compared three classifiers—XGBoost (recall 0.36), Random Forest (recall 0.48; 5‑fold CV mean recall 0.48 ± 0.02), and a soft voting ensemble (recall 0.45). The most influential predictors—MonthlyIncome, Age, and JobSatisfaction—emerged consistently. I recommend targeted retention efforts for mid‑career, higher‑paid employees showing low job satisfaction.

**Baseline Establishment**

I calculated the overall attrition rate to set a point of comparison for my models and business impact assessment:

**Baseline attrition:** 29.60%  
**

A nearly 30 % turnover rate represents a substantial cost in recruitment, training, and lost productivity.

**Exploratory Data Analysis (EDA)**

**Monthly Income & Attrition**

A boxplot reveals that employees who left tend to have slightly higher median income but also greater dispersion (outliers at both ends):  
*A graph showing a comparison of income and attrition

AI-generated content may be incorrect.*

**Insight:** Compensation alone does not guarantee retention—higher‑paid employees may leave if other factors (e.g., career growth) are lacking.

**Attrition by Department**

The countplot shows Research & Development and Sales have the highest churn counts:  
*A graph of a bar graph

AI-generated content may be incorrect.*

**Insight:** Departments with demanding targets or project cycles (R&D, Sales) may require focused retention strategies.

**Job Role & Income Trends**

Grouping rare roles into “Other,” I compared income distributions by role and attrition status:  
*A graph of red and blue squares

AI-generated content may be incorrect.*

**Insight:** Sales Executives and Research Directors who leave earn more on average, suggesting that top performers may be more mobile.

**Job Satisfaction Interaction**

Overlaying JobSatisfaction reveals that low satisfaction (1–2) aligns with lower incomes but still sees attrition:  
*A graph of different colored and black rectangles

AI-generated content may be incorrect.*

**Insight:** Even modest income increases don’t fully offset dissatisfaction; career fulfillment initiatives are critical.

**Correlation Structure**  
*A chart of numbers and symbols

AI-generated content may be incorrect.*

The heatmap displays Pearson correlations among all numeric variables, with the Attrition column revealing each feature’s linear relationship to turnover.

**Implication:** Although pay and tenure variables track closely with each other, their straight‑line correlations with Attrition are minimal. This underscores why we leverage non‑linear, interaction‑aware models (e.g. Random Forest) to capture the complex drivers of employee turnover.

**Insight:** Attrition is a multifactorial phenomenon; no single feature dominates, necessitating multivariate models.

**Feature Engineering**

To prepare for predictive modeling, I:

**Encoded OverTime** (Yes→1, No→0)

**Grouped JobRole** into “Other” if count < 5 to reduce sparsity

**One‑hot encoded** BusinessTravel, Department, and JobRole\_GroupedA screenshot of a computer

AI-generated content may be incorrect.

I selected as model inputs:

**Numeric:** Age, MonthlyIncome, JobSatisfaction, OverTime\_Encoded

**Categorical dummies:** BusinessTravel\_*, Department\_*, JobRole\_\*

**Predictive Modeling & Evaluation**

**Data Preparation Pipeline**

I built a pipeline to impute missing values (median for numeric, most frequent for categorical), encode categoricals, and apply SMOTE (Maklin, 2022) on the training set to correct class imbalance.

**Model Benchmarks**

**K – Fold & Model Accuracy** (Brownlee, 2023)

A black background with white numbers

AI-generated content may be incorrect.

We validated our tuned Random Forest with 5‑fold cross‑validation, achieving a mean recall of 0.483 (σ = 0.024), consistent with our hold‑out recall of 0.481. This demonstrates stable, reliable performance across different data partitions.

**XGBoost Classifier** (geeksforgeeks, 2025)

Recall: 0.36

Accuracy: 0.74

**Random Forest (GridSearchCV)** (scikit-learn, 2025)(geeksforgeeks, 2025)

Best params: n\_estimators=100, max\_depth=None, min\_samples\_split=2, min\_samples\_leaf=1

Recall: 0.48

Accuracy: 0.77

*A chart of different colors

AI-generated content may be incorrect.*

**Soft‑Voting Ensemble** (LR + RF + XGB)

Recall: 0.45

Accuracy: 0.76

| **Model** | **Accuracy** | **Precision** | **Recall** | **F₁‑Score** |
| --- | --- | --- | --- | --- |
| XGBoost | 0.74 | 0.55 | 0.36 | 0.44 |
| RandomForest | 0.77 | 0.61 | 0.48 | 0.54 |
| Ensemble | 0.76 | 0.60 | 0.45 | 0.52 |

**Conclusion:** The Random Forest strikes the best balance of recall and overall accuracy.

**Feature Importance & SHAP Analysis** (Scott M. Lundberg, 2017)

The Random Forest’s built‑in feature importances highlight:

A graph with red and blue bars

AI-generated content may be incorrect.

| **Rank** | **Feature** | **Importance** |
| --- | --- | --- |
| 1 | MonthlyIncome | 0.342 |
| 2 | Age | 0.246 |
| 3 | JobSatisfaction | 0.120 |

To explain nonlinear effects, I used SHAP (Lundberg & Lee, 2017):

**Top 10 SHAP bar plot** confirms MonthlyIncome and Age are dominant drivers.  
*A blue rectangular object with white text

AI-generated content may be incorrect.*

I further experimented with **top 5** and **top 3** features to demonstrate compute trade‑offs:

A screen shot of a computer

AI-generated content may be incorrect.

**Top 5 beeswarm** (5 features): fast (< 0.04 s)

**Top 3 beeswarm** (3 features): slower but interpretable (~ 8 s)

SHAP timings illustrate that bar charts deliver nearly instant insight, while deep beeswarm plots may require more computing time.

**Conclusions & Recommendations**

**Why is attrition high?**

Mid‑career employees with higher incomes are most at risk, likely to seek new challenges or rewards.

Departments with intense workloads (R&D, Sales) exhibit greater churn.

**How can we address it?**

**Targeted retention programs** for the “MonthlyIncome–Age” cohort: mentorship, career development plans, stretch assignments.

**Adjust compensation frameworks** to better reward longevity and performance.

**Monitor job satisfaction** via pulse surveys and act on low‑satisfaction flags.

**Next Steps**

Deploy the Random Forest model into a quarterly dashboard for HR.

Extending analysis with text mining on exit interviews.

Conduct A/B tests on flexible work pilots.

By combining robust EDA, feature engineering, and interpretable machine learning, I have determined actionable insights and built a predictive tool that your leadership team can use to reduce turnover and support organizational success.

# References

Brownlee, J. (2023, Oct 4). *A Gentle Introduction to k-fold Cross-Validation*. Retrieved from machinelearningmastery: https://machinelearningmastery.com/k-fold-cross-validation/

geeksforgeeks. (2025, Feb 12). *ML | XGBoost (eXtreme Gradient Boosting)*. Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/ml-xgboost-extreme-gradient-boosting/

geeksforgeeks. (2025, Jan 16). *Random Forest Algorithm in Machine Learning*. Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

Jang, D. (2024, Oct 20). *Exploring the Scikit-Learn Classification API: A Deep Dive into Supervised Learning Techniques*. Retrieved from medium: https://medium.com/@jangdaehan1/exploring-the-scikit-learn-classification-api-a-deep-dive-into-supervised-learning-techniques-3dc1c404677d

Maklin, C. (2022, May 14). *Synthetic Minority Over-sampling TEchnique (SMOTE)*. Retrieved from medium: https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c

matplotlib. (2025). *Matplotlib: Visualization with Python*. Retrieved from matplotlib: https://matplotlib.org/

scikit-learn. (2025). *6.3. Preprocessing data*. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/preprocessing.html

scikit-learn. (2025). *6.4. Imputation of missing values*. Retrieved from scikit-learn.: https://scikit-learn.org/stable/modules/impute.html

scikit-learn. (2025). *GridSearchCV*. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html

scikit-learn. (2025). *sklearn.metrics*. Retrieved from scikit-learn: https://scikit-learn.org/stable/api/sklearn.metrics.html

Scott M. Lundberg, S.-I. L. (2017). A Uniﬁed Approach to Interpreting ModelPredictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768 - 4777). Red Hook, NY, United States: Curran Associates Inc. Retrieved from dl.acm.

seaborn. (2025). *seaborn: statistical data visualization*. Retrieved from seaborn: https://seaborn.pydata.org/index.html