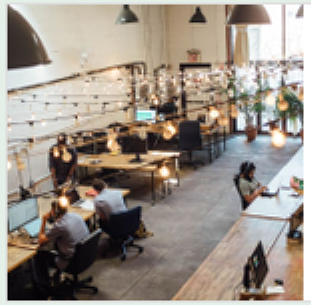
HR ANALYTICS DATASET

SHLOMI SHOR III



HR Analytics Dataset

Unlocking HR Insights: Data-driven Decisions

kaggle.com

https://www.kaggle.com/datasets/saadharoon27/hr-analytics-dataset

The task is to analyze the target column **"Attrition"** (which indicates whether an employee left the company or not), and using models to predict the likelihood of leaving.

DATA OVERVIEW (EDA)

Dataset Structure

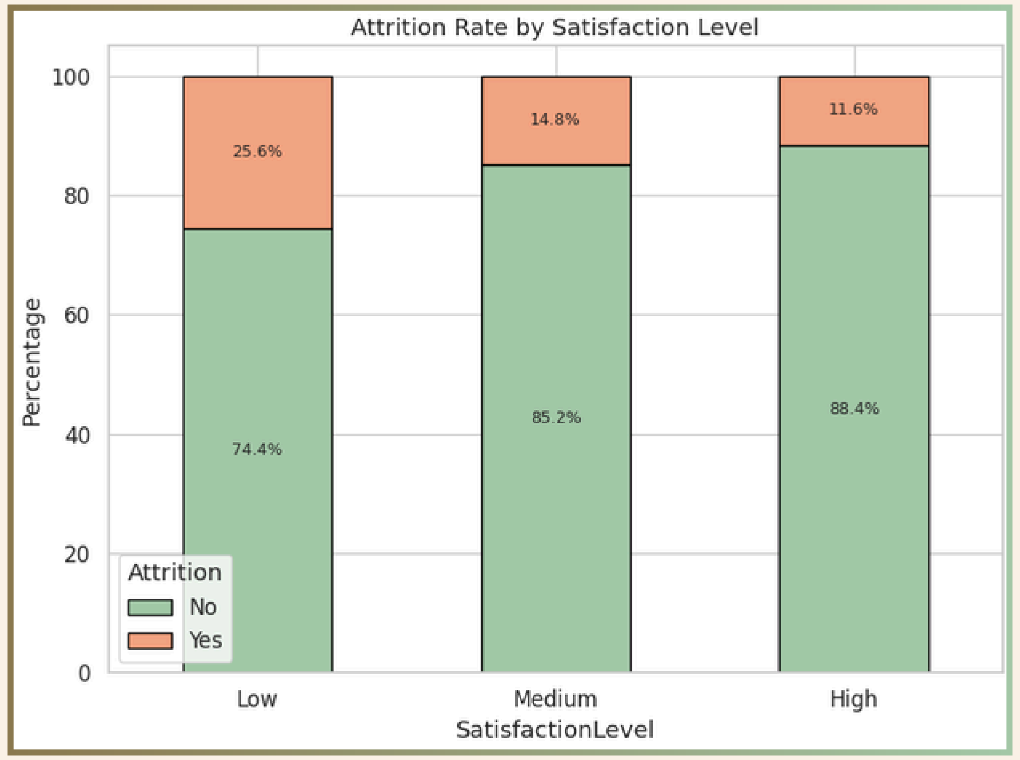
- **1470** rows and **35** columns.
- **9** Categorial columns.
- **26** Numerical columns.

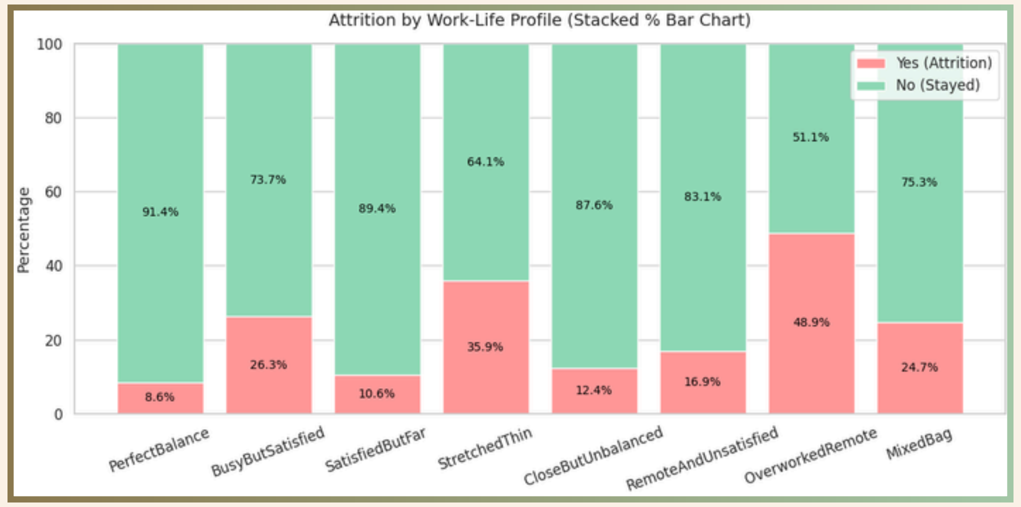
Data Quality and Cleaning

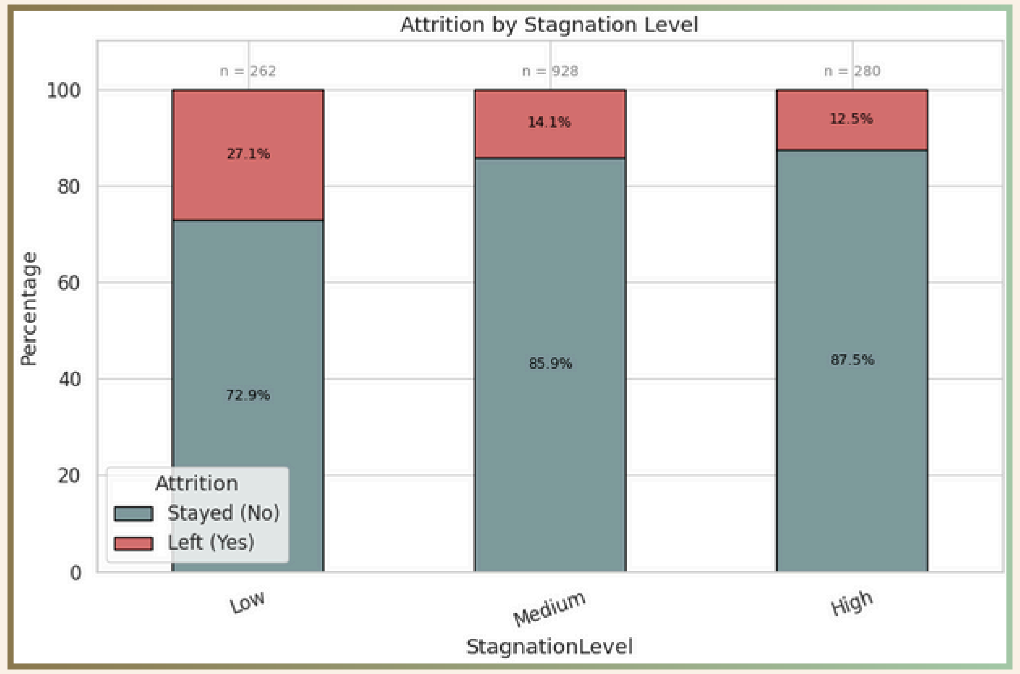
- During the data quality checks, we found no duplicates, no negative values in key fields, no zero monthly incomes, and no logical inconsistencies between time and experience fields.
- **We removed four non-informative columns**: three with constant values (EmployeeCount, Over18, StandardHours) and one ID column (EmployeeNumber).
- We suspected **Performance Rating** and **Percent Salary Hike** might cause **Leackge** by accidentally signaling who's leaving, as sometimes these metrics only track active employees. However, after checking, we found these columns are **available for all employees**, making them reliable and safe for our model to use.

Attrition

- only **16.1%** (237 employees) **left** the company, while **83.9%** (1,233 employees) **stayed**.
- We examined how **satisfaction impacts employee attrition by combining three measures** – job satisfaction (JobSatisfaction), environment satisfaction (EnvironmentSatisfaction), and relationship satisfaction (RelationshipSatisfaction) – into one overall indicator. We found that when overall satisfaction is high, the attrition rate significantly decreases.
- We examined how **workload, work-life balance, and commute distance** impact employee attrition by combining these three variables into distinct work-life profiles. We found that **employees with balanced conditions are significantly less likely to leave, while those experiencing both overtime and poor balance show the highest attrition rates**.
- We examined how professional stagnation impacts employee attrition by combining **four factors into a stagnation score**: years since last promotion (YearsSinceLastPromotion), years in current role (YearsInCurrentRole), job level (JobLevel), and years with the current manager (YearsWithCurrManager). We found that employees with **higher stagnation scores**, **those more professionally "stuck"**, are significantly **less likely** to **leave** (**12.5%**), while those with **lower stagnation scores** have much **higher attrition rates** (**27.1%**).







MODEL 1 + MODEL 2

Our Approach & Key Decisions -

- Our **core aim** was to **spot employees at risk of leaving as early as possible**. We focused on '**Recall**' because it's better to have a **few false alarms** than to miss someone who actually departs.
- We compared various **data balancing methods**, and **Oversampling alone** proved most effective for **Recall**. This allowed our model to **learn more accurately** from the actual cases of employees who left.
- We removed the variables [**'EmployeeNumber', 'EmployeeCount', 'Over18', 'StandardHours'**] before running the model because they **don't contribute to predictions** and could even **harm the model's quality**.
- In the second model, we included **two engineered features WorkLifeProfile** and **SatisfactionLevel**, as part of the ten selected variables. The remaining eight variables were chosen based on their **strong correlation** with **Attrition**.

Full DataBase Logistic Regression

	Predicted: No	Predicted: Yes
Actual: No	272 (61.7%)	99 (26.7%)
Actual: Yes	17 (3.9%)	53 (12%)

Accuracy: 0.737 **Recall:** 0.757

Precision: 0.349 **F1 Score:** 0.477

10 Variable Logistic Regression

YearsWithCurrManager	Gender
JobLeve	YearsInCurrentRole
Age	MaritalStatus
SatisfactionLevel	WorkLifeProfile
YearsSinceLastPromotion	MonthlyIncome

	Predicted: No	Predicted: Yes
Actual: No	283 (64.2%)	88 (20%)
Actual: Yes	18 (4.1%)	52 (11.8%)

Accuracy: 0.759 **Recall:** 0.743

Precision: 0.371 **F1 Score:** 0.495

MODEL 1 + MODEL 2

Results -

- Our **refined model**, using just **10 key variables**, showed overall **better performance** with higher **Accuracy (76%)** and **Precision (37.1%)**, while maintaining strong **Recall (74.3%)**.
- The improved performance, despite fewer features, highlights the power of our newly created features: **WorkLifeProfile** and **SatisfactionLevel**. They provide a **more accurate read on the employee experience**, leading to **more precise predictions**.
- Crucially, our model became **smarter about false alarms**, reducing **false positives from 99 to 88**. While Recall had a **slight, acceptable dip (from 75.7% to 74.3%)**, the number of **actual leavers identified remained almost the same** across models (**53 vs. 52**).
- The second model is more operationally efficient, providing greater accuracy with fewer variables.

Model 1 vs. Model 2: Top Attrition Drivers -

- **Model 1:** Focuses on **Overtime**, **Stock Option Level**, and **Education Field (Medical)**, expressing the impact of **existing operational and demographic characteristics** on employee attrition.
- **Model 2:** The three most significant features are based on our **engineered features**. The focus here is on the employee's **internal and emotional experience**, such such as **Satisfaction Level** and **Work-Life Balance**.
- Ultimately, while **Model 1** might suggest retention strategies focused on **external motivation** (like **pay** and **bonuses**), **Model 2's refined approach** shows us something deeper: that **internal motivation**, stemming from **employee happiness**, **work-life balance**, and **good relationships** in the company, truly makes people **stay**.

Model 1	
OverTime_Yes	+0.635
StockOptionLevel	−0.348
EducationField_Medical	−0.322

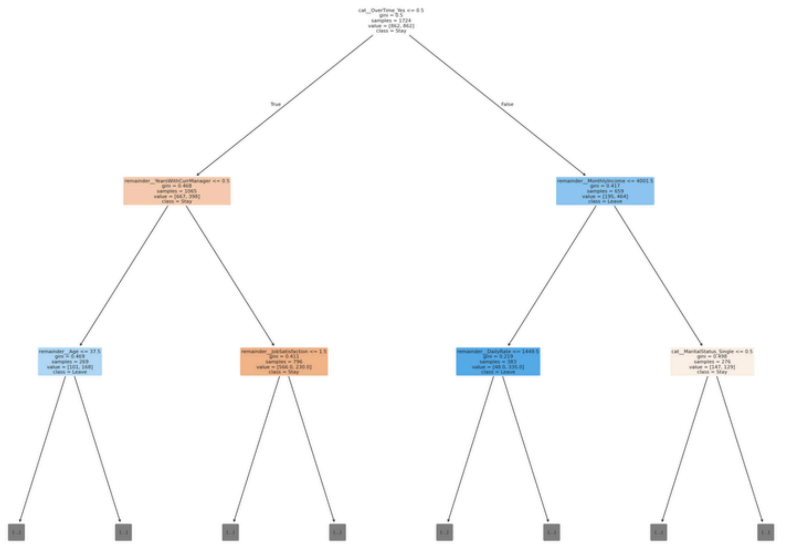
Model 2	
SatisfactionLevel_Low	+1.709
WorkLifeProfile_PerfectBalance	−1.479
WorkLifeProfile_SatisfiedButFar	−1.282

DECISION TREE VS. LOGISTIC REGRESSION MODELS

Decision Tree

	Predicted: No	Predicted: Yes
Actual: No	328 (74.4%)	43 (9.7%)
Actual: Yes	49 (11.1%)	21 (4.8%)

Accuracy: 0.791 **Precision:** 0.328 **Recall:** 0.3



Comparison between models

Metric	Model 1	Model 2	Decision Tree
Accuracy	0.737	0.760	**0.791**
Precision	0.349	**0.371**	0.328
Recall	**0.757**	0.743	0.3
F1 Score	0.477	**0.495**	0.313

Insights -

- **Accuracy** - The Decision Tree achieved a **high Accuracy of 79.1%** , but it missed **over two-thirds** of the actual employees who left (**low Recall of 30.0%**). Therefore, **Accuracy alone is not enough** for our goal of identifying employees **before they leave**.
- **Recall** - Our defined goal is to **identify employees before they leave**. For this reason, **models with high Recall**, such as **Model 1 (75.7%)** and **Model 2 (74.3%)**, are **significantly preferred** over the Decision Tree (**30.0%**).
- **Model 2 Offers the Best Balance -** It maintains **high Accuracy (76.0%)** and **Recall (74.3%)**, and achieved the **highest F1-SCORE (0.495)** (which is the harmonic mean of Precision and Recall). Therefore, **among the three models**, **Model 2 is currently the most optimal**.

TUNED DECISION TREE

We trained two tuned trees using different feature sets -

- **Tuned Tree (Model 1) -** All original features (after encoding).
- **Tuned Tree (Model 2) -** 10 selected features, including engineered ones like SatisfactionLevel and WorkLifeProfile.
- Both trees were trained using the same tuning parameters to reduce overfitting and improve generalization: max_depth = 5, min_samples_split = 10, min_samples_leaf = 4.

Insights -

- The **unbounded Decision Tree (B4)** had a very low Recall of just **30%**. However, when we **constrained the tree (B5)**, its Recall increased to **47%**. This means we significantly improved the model's ability to **identify employees likely to leave**.
- The **refined Decision Tree with engineered features** achieved an **F1 Score** of **36.2%**, and its Recall also improved to **45.7%** compared to the unbounded tree. This shows that our new features helped the model identify more employees who actually left. This highlights that **good feature selection is just as important as the model itself**.
- Despite improvements in the tuned Decision Trees, Logistic Regression with engineered features (**Model 2**) still delivered the **best results,** with the **highest F1 Score (49.5%)** and **Recall (74.3%)**. This shows that **simpler, regularized models can outperform complex ones** when paired with **smart features**.

Comparison between models

Metric	Unbounded Tree (B4)	Tuned Tree (Model 1)	Tuned Tree (Model 2)
Accuracy	**0.791**	0.678	0.744
Precision	**0.328**	0.239	0.299
Recall	0.3	**0.471**	0.457
F1 Score	0.313	0.317	**0.362**

Model 1

	Predicted: No	Predicted: Yes
Actual: No	266 (60.3%)	105 (23.8%)
Actual: Yes	37 (8.4%)	33 (7.5%)

Model 2

	Predicted: No	Predicted: Yes
Actual: No	296 (67.1%)	75 (17%)
Actual: Yes	38 (8.62%)	32 (7.26%)

MODEL 2 WINS AND IT MATCHES WHAT THE EDA TOLD US

The best model is Logistic Regression (Model 2) -

- Its **Recall** is **very high (74.3%)**, meaning the model identifies nearly three-quarters of actual leavers. This is precisely what the business needs to intervene in time.
- It has the **highest F1 Score (0.495)**, indicating the model not only identifies leavers but also maintains a strong balance between Recall and Precision.
- The model is built on **10 explainable and monitorable features** that can be measured and tracked over time, making it highly applicable in the real world.
- **Bottom Line:** Logistic Regression with engineered features provides the business with the clarity, accuracy, and actionable insights it needs to reduce attrition risk.

Alignment Between EDA and Model: Key Insights -

- **Satisfaction:** In our Exploratory Data Analysis (EDA), we found a **significant link** between overall satisfaction (a combination of Job, Environment, and Relationship satisfaction) and the tendency to leave. Employees with **high satisfaction almost never leave**. In Model 2, the engineered feature **SatisfactionLevel_Low** (+1.71) was the **strongest predictor**, showing **full alignment** between our EDA insights and the model's results.
- **High Work-Life Balance:** Our EDA revealed that employees with a **high work-life balance** (no overtime, short commute, high work-life balance) are l**ess likely to leave**, while those with a combination of high workload, low balance, and long commute are most likely to leave. In Model 2, the engineered features **WorkLifeProfile_PerfectBalance** (-1.48) and **WorkLifeProfile_SatisfiedButFar** (-1.28) were among the **three strongest variables**. Here too, there is **full alignment** between the EDA findings and the model's results.
- **Professional Stagnation:** In our EDA, we discovered that employees with a **high feeling of stagnation** leave less often (only 12.5%), while those **without a path for advancement** leave more often (27.1%). However, no variable directly related to professional stagnation emerged as a dominant predictor in our models. This suggests a **more complex impact** or that **deeper analysis is needed** for this factor.
- **Monthly Income:** Surprisingly, in the EDA phase, **monthly income was not identified as a significant variable** in employee attrition. However, in **all three Decision Trees** (including the unbounded B4 and the tuned trees), it appeared among the **three strongest features**. This was an insight that did not emerge from the EDA but proved important in the modeling phase, highlighting the **value of combining initial intuition with algorithmic analysis**.

BUSINESS INSIGHTS AND CONCLUSION

Business Insights & Recommendat

- **Good Work-Life Balance Significantly Reduces Attrition Risk -** Our analysis shows that a **"PerfectBalance"** profile is a **strong protective factor** against attrition. **Conversely**, employees in "**OverworkedRemote"** profiles face **high risk**.
  - **Business Recommendation:** Proactively foster a culture of work-life balance by enabling flexible work arrangements (remote work, flexible hours) and providing robust technological support. This is particularly vital for employees with high workloads or long commutes, as it directly mitigates their attrition risk.
- **Unhappy Employees Leave - Satisfaction is a Strong Indicator**, The engineered feature **SatisfactionLevel_Low (+1.71)** was the **strongest variable** in **Model 2** and also appeared in **other models**.
  - **Business Recommendation:** Enable remote work, flexible hours, and technological support, especially for employees with high workloads or long commutes. By fostering a better work-life balance, we directly reduce attrition risk.
- **Monthly Income: A Consistent Predictor in Decision Trees -** While not prominent in our initial **EDA, MonthlyIncome** consistently appeared among the **top three features** in all **Decision Tree models** (both unbounded and refined).
  - **Business Recommendation: Check employee salary** alignment with market average semi-annually, and consider **upgrading compensation** for employees **below the industry threshold**. Ensuring fair and competitive pay is a fundamental retention strategy.
- **Feeling of Stagnation Doesn't Always Lead to Leaving -** Our **EDA** suggested that **professionally stagnant employees** might leave **less** than those with advancement paths. **However**, related features (e.g., **YearsInCurrentRole**, **YearsSinceLastPromotion**) were **not dominant predictors** in our models.
  - **Business Recommendation: Focus retention efforts** on **new and mid-tenure employees** through **onboarding conversations, satisfaction monitoring**, and **rapid development programs**. Long-tenured employees are **generally not at risk** of leaving, and thus **do not require immediate intervention**.

Conclusion

During our work, we strategically combined **Exploratory Data Analysis (EDA) with model building** to predict employee attrition. We found a **significant overlap** between the two approaches, especially in identifying factors like **employee satisfaction and work-life balance**, which were confirmed by both methods. However, the EDA highlighted certain variables (such as "professional stagnation") that did not emerge as dominant factors in the models, while the models revealed the importance of other variables (like "monthly income") that weren't prominent in the initial EDA. This combination of approaches allowed us to build a **holistic and data-driven picture** of attrition drivers. Therefore, we **highly recommend full integration of EDA and model building in every data analysis project**.