EMPLOYEE ATTRITION

BUSINESS CASE

Employee attrition remains a challenge at any company. The departure of skilled employees disrupts innovation, affects project timelines, and adds substantial costs related to knowledge loss, recruitment, and training. This extensive HR dataset created by IBM Datascientists hold valuable insights into factors influencing attrition, but a focused analytical approach is needed to turn this data into actionable retention strategies.

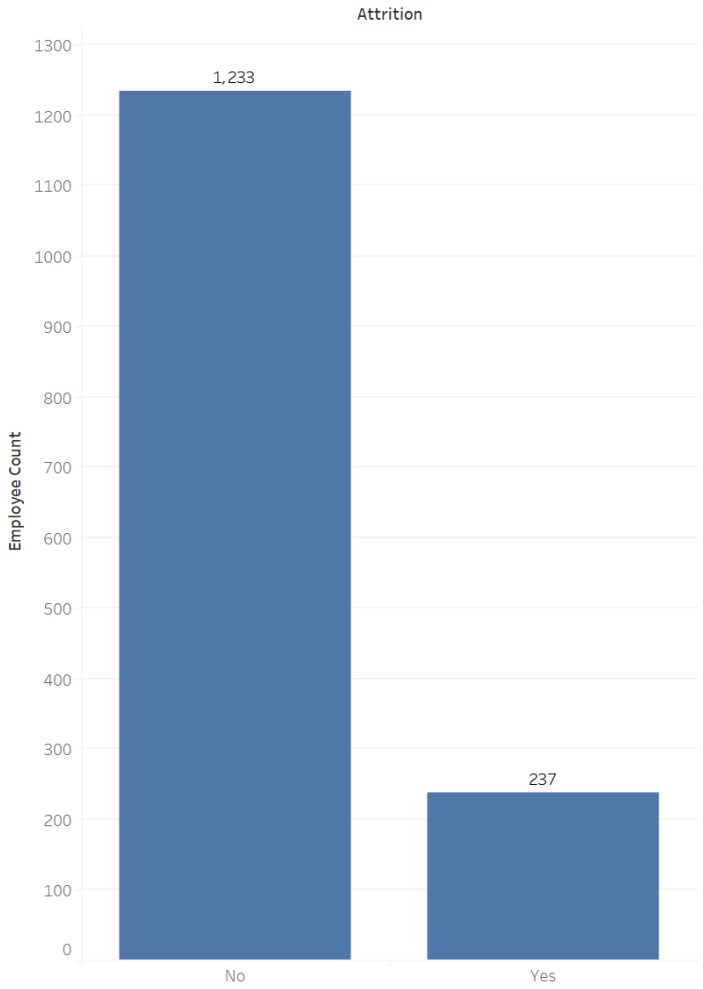
By implementing machine learning models to predict employee attrition, we can proactively identify individuals at high risk of leaving. This proactive approach will enable targeted retention strategies, minimizing the financial burden of turnover, ensuring workforce stability, and protecting company's valuable intellectual capital. Prioritizing employee retention will not only reduce costs but also enhance morale, fostering a work environment that champions employee satisfaction and loyalty.

For this business case, the most appropriate metric for evaluating model performance is **F1-Score** because:

- Unlike accuracy, F1-Score considers both precision and recall, which is crucial in our case. Since there are likely far fewer employees leaving than staying, a model with high accuracy might simply be good at predicting the majority class (staying). F1-Score balances these factors, giving us a more accurate picture of the model's ability to identify both high and low-risk employees.
- F1-Score considers the cost of misclassifications. A false positive (predicting someone will leave when they stay) wastes resources, but a false negative (missing an at-risk employee) has far greater consequences. F1-Score strikes a balance between these two scenarios.

EXPLORATORY DATA ANALYSIS

- There are 1470 rows of employee data with 35 columns/features (including Attrition). There are 26 numerical features and 9 categorical features (including Attrition).
- We will also check for any class imbalance in our target variable, 'Attrition'.

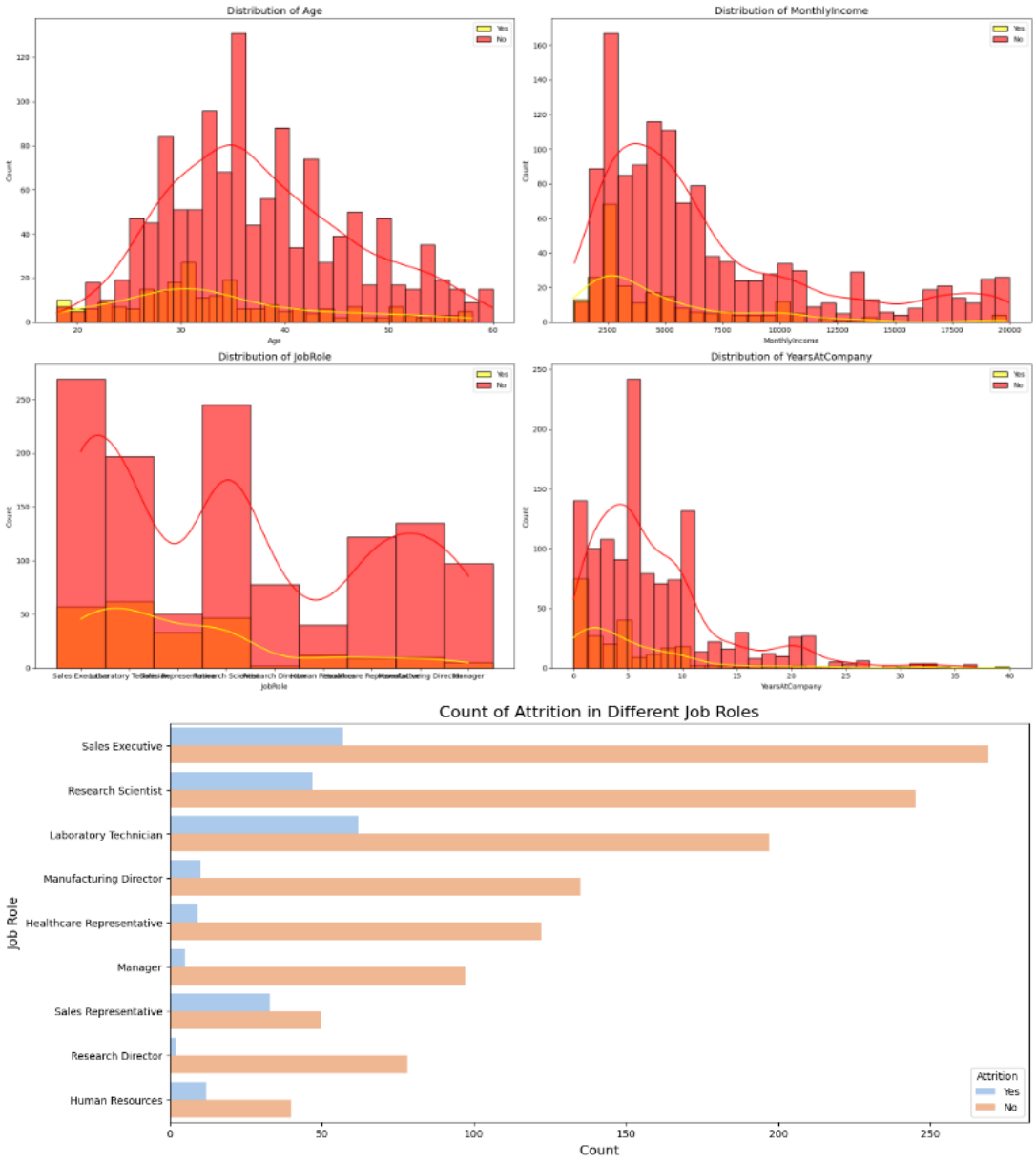


The 'Attrition' column shows a significant class imbalance with 1233 'No' and 237 'Yes'.

Distribution of Key Features by Attrition

As we can see from the figure below (coded in Python) that

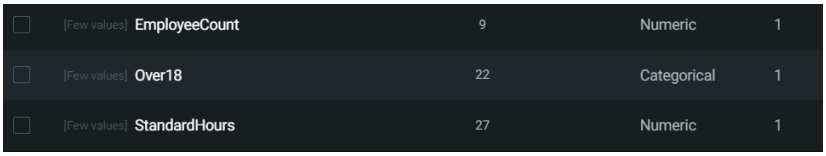
- The data is inclined towards research-backed evidence.
- There is a noticeable difference between employees who left and those who stayed for features like Age, Monthly Income and Years at Company.
- Some job roles experience more attrition (ratio wise) than other positions - notably with Sales Executive, Research Scientist, Laboratory Technician, Sales Representative - and less in more management roles - notably Director and Manager level job roles.



DATA PRE-PROCESSING

There are no missing values in the dataset.

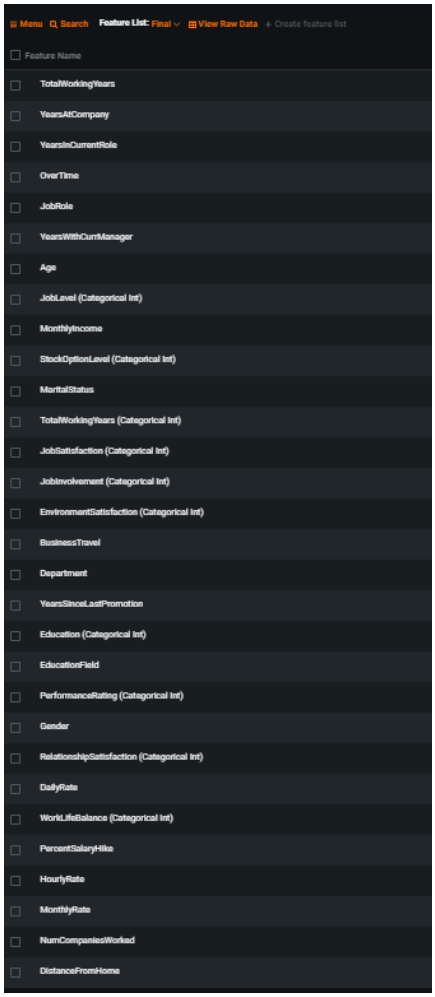
While there are no missing values found, 3 features ('EmployeeCount', 'Over18', 'StandardHours') had 1 unique value. Also, 'EmployeeNumber' is a unique identifier. Thus, when creating a new feature list “Final”, these 4 features have been dropped as they would not be helpful in predicting attrition.



Furthermore, following pre-processing was done on features through DataRobot.

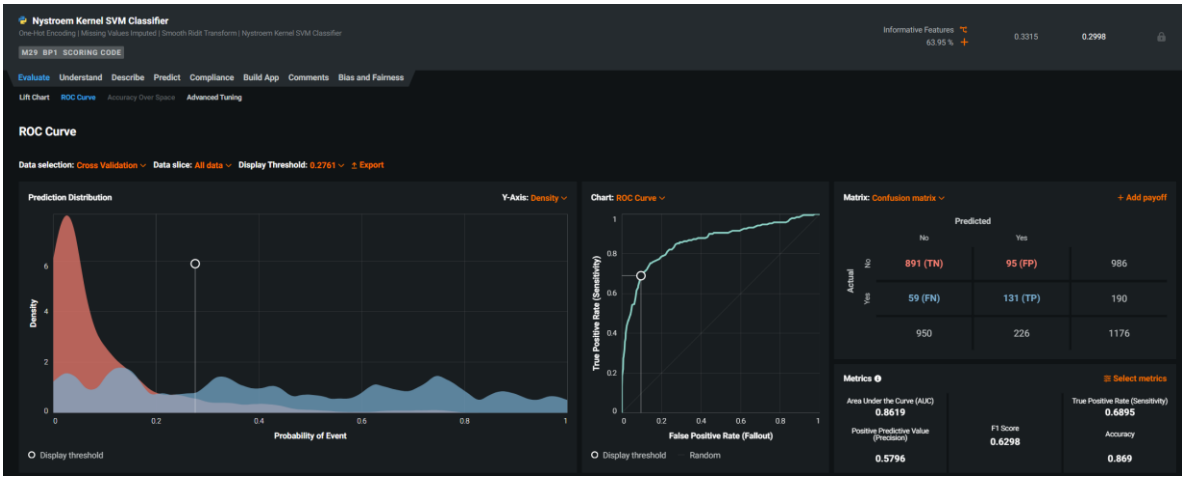
1. Converted **StockOptionLevel** from *Numeric* to *Categorical* data
2. Converted **JobLevel** from *Numeric* to *Categorical* data
3. Converted **JobInvolvement** from *Numeric* to *Categorical* data
4. Converted **JobSatisfaction** from *Numeric* to *Categorical* data
5. Converted **EnvironmentSatisfaction** from *Numeric* to *Categorical* data
6. Converted **JobSatisfaction** from *Numeric* to *Categorical* data
7. Converted **Education** from *Numeric* to *Categorical* data
8. Converted **PerformanceRating** from *Numeric* to *Categorical* data
9. Converted **RelationshipSatisfaction** from *Numeric* to *Categorical* data
10. Converted **WorkLifeBalance** from *Numeric* to *Categorical* data
11. Converted **TotalWorkingYears** from *Numeric* to *Categorical* data

The “Final” feature list was then created:

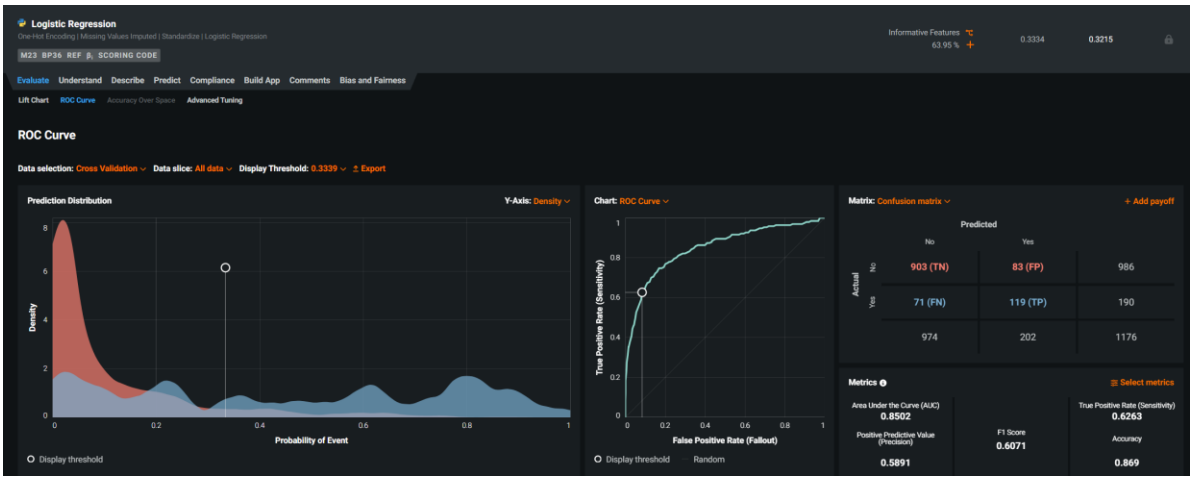


PREDICTION MODELS

SVM Classifier:

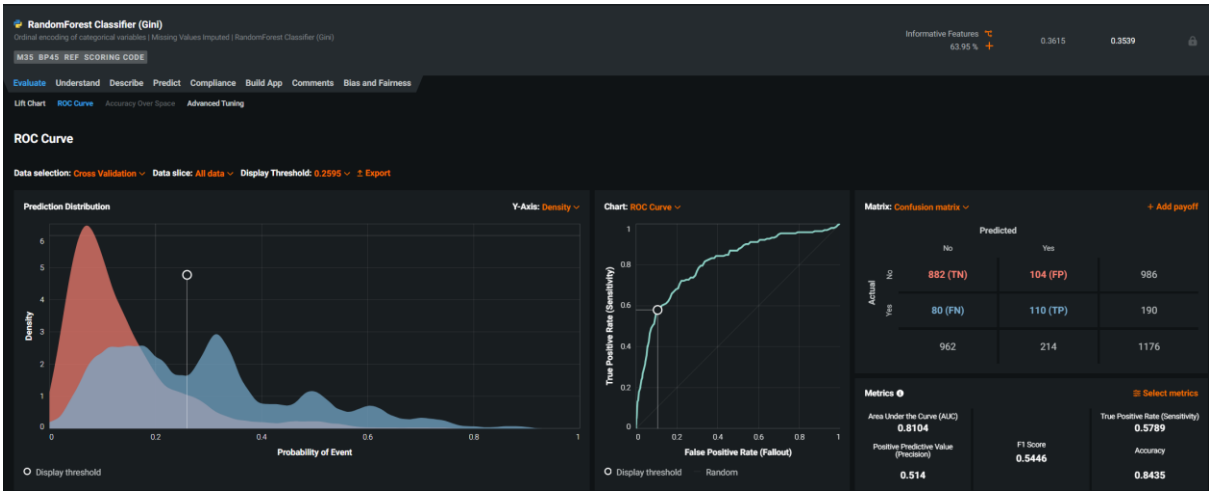


Logistic Regression:



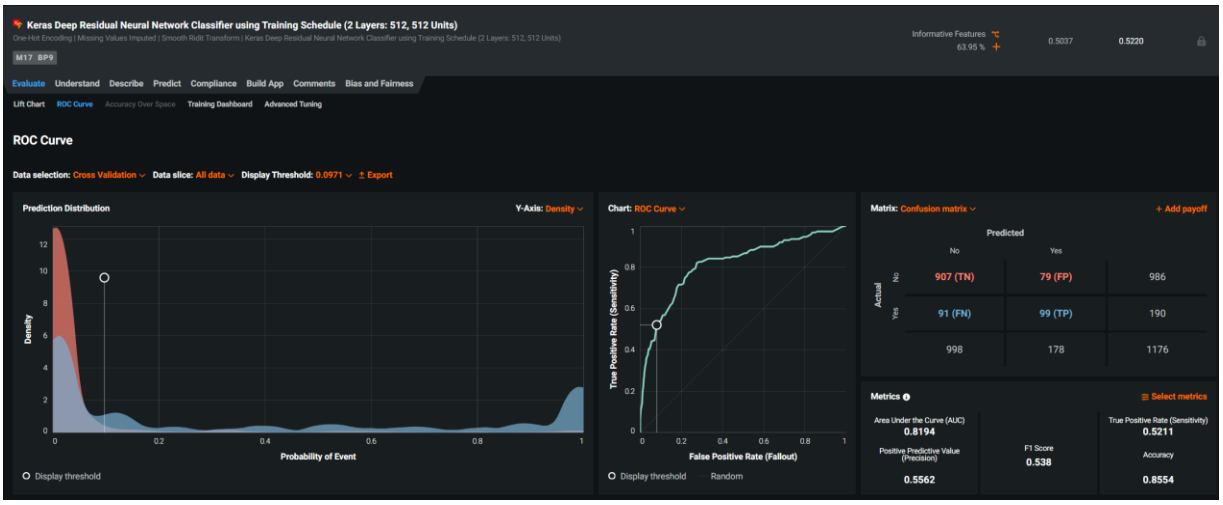
Boosted Trees:

Random Forest:



K-Nearest Neighbors Classifier:

Neural Network CLaasifier:



EVALUATION OF MODELS

For the given business case, this is the meaning of each metrics:

- Accuracy: This is the most basic metric, and it simply measures the percentage of predictions that are correct. However, accuracy can be misleading in imbalanced datasets, where there are significantly more employees who stay with the company than employees who leave.
- Precision: This metric measures the proportion of employees predicted to leave who actually leave. It is important to consider precision if the cost of misclassifying an employee as a flight risk is high.
- Recall: This metric measures the proportion of employees who actually leave who are correctly predicted to leave. It is important to consider recall if the cost of failing to identify a flight risk is high.
- F1-score: This metric is a harmonic mean of precision and recall, and it provides a balance between the two metrics.
- Area Under the ROC Curve (AUC): The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for different classification thresholds. The AUC is a measure of the overall performance of a model at classifying between employees who leave and employees who stay.

We cannot particularly calculate the payoff matrix because

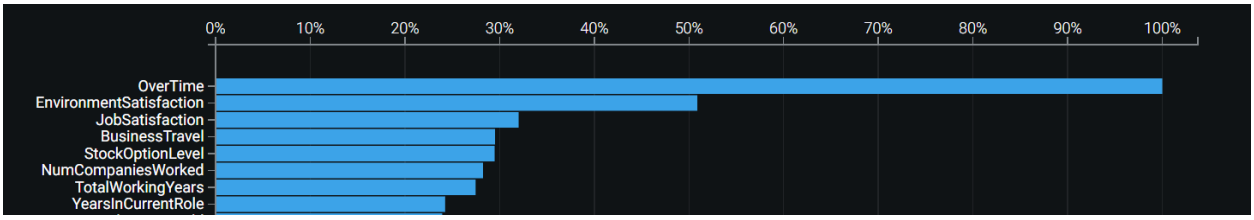
- The salary varies from person to person.
- The revenue earned by the company varies from person to person.
- Employee attrition prediction is more about probabilities than fixed outcomes. One wants a model that can reliably provide the likelihood an employee is considering leaving. This allows for more nuanced business decisions than simply classifying employees into a 'will leave' or 'will stay' bucket with arbitrary payoffs attached.
- In this dataset, we have far more instances of employees leaving the company than those leaving. This makes a standard payoff matrix unreliable as the focus becomes heavily skewed towards accurately predicting the majority class (employees staying).

SUMMARY OF EACH MODEL

	SVM Classifier	Logistic Regression	Boosted Trees	Random Forest	KNN Classifier	Neural Network Classifier
Accuracy	0.869	0.869	0.827	0.844	0.821	0.855
Precision	0.58	0.59	0.474	0.514	0.455	0.556
Sensitivity	0.69	0.626	0.621	0.579	0.553	0.521
F1 Score	0.63	0.607	0.538	0.545	0.499	0.538
AUC of ROC	0.862	0.85	0.806	0.81	0.778	0.819

Based on the summary and the metrics discussed in the business case, the best performing model is **SVM Classifier**.

FEATURE IMPACT AND RECOMMENDATIONS



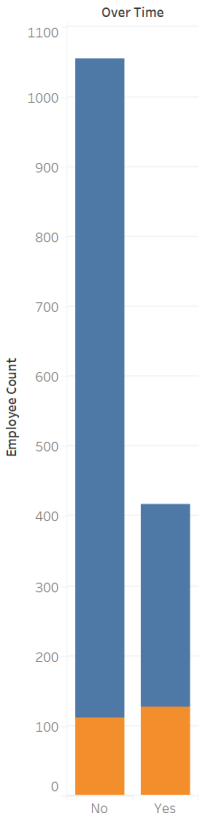
The top 5 features with the observations and actionable recommendations if any are as follows:

.1. Overtime:

Employees working excessive overtime are more likely to experience burnout and dissatisfaction, leading to increased attrition risk. We can see that when we compare the ratios of employees leaving to employees staying

Actionable recommendations:

- Analyze workload distribution and identify areas where overtime can be reduced or better compensated.
- Implement flexible work arrangements where possible to help employees with work-life balance.
- Promote a culture that values well-being and discourages excessive overtime.

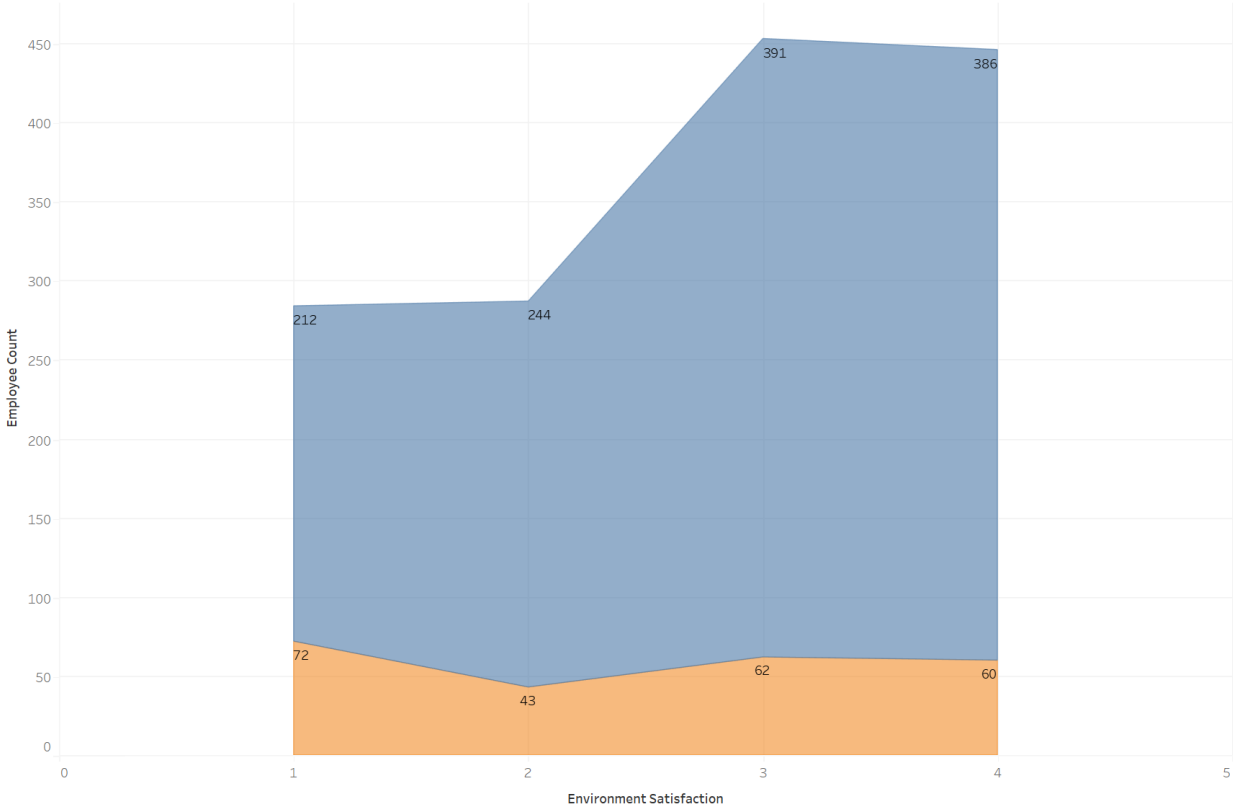


.2. Environment Satisfaction:

Employees with 'Low' or 'Medium' environment satisfaction have an increased risk of leaving (comparing the ratios of employees leaving to employees staying). Those with 'High' or 'Very High' satisfaction are more likely to stay.

<u>Actionable recommendations:</u>

- Conduct surveys with targeted questions to understand the specific aspects driving dissatisfaction in the work environment.
- Invest in improvements to physical workspace, collaboration tools, or aspects employees value.
- Foster a positive team culture through team-building activities and recognition programs.

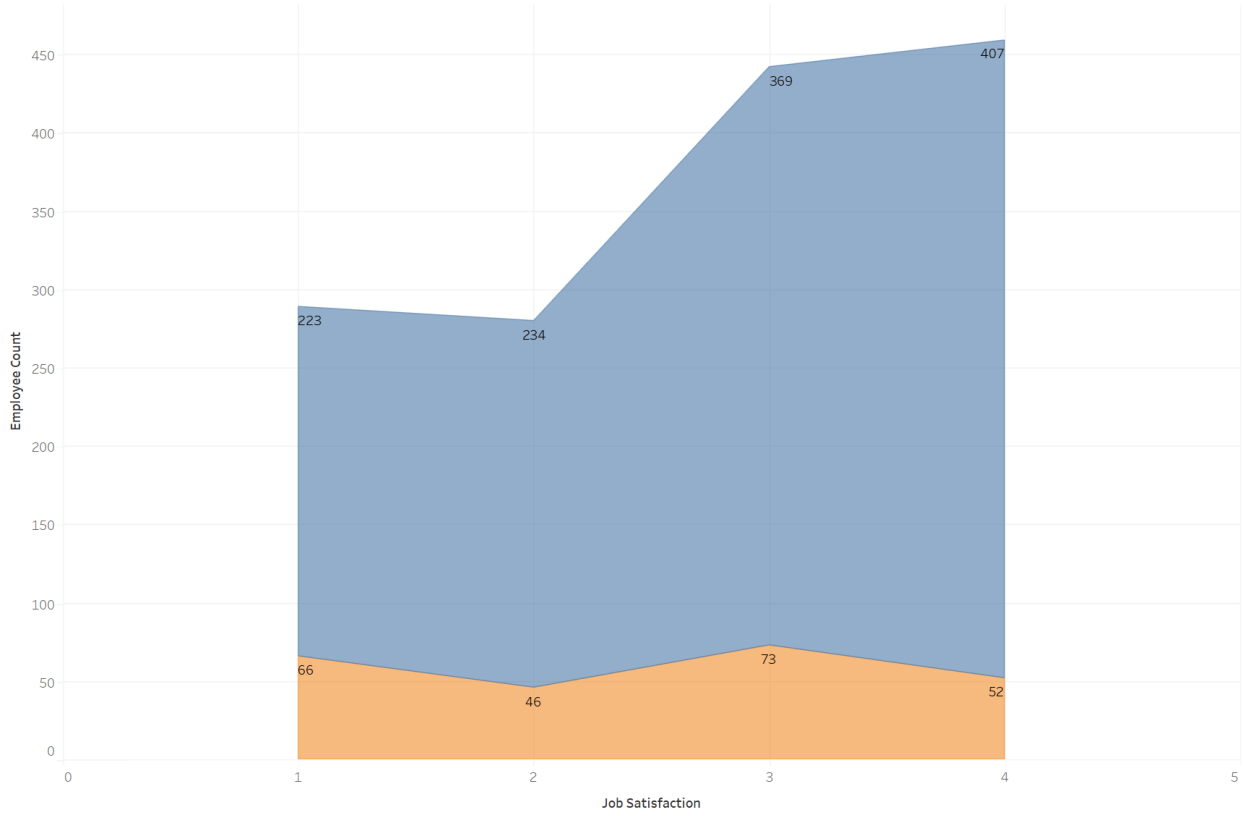


.3. Job Satisfaction:

Employees with 'Low' or 'Medium' job satisfaction are more likely to leave, which is clearly visible in the following visualisation. Those with 'High' or 'Very High' satisfaction exhibit greater loyalty and retention potential.

Actionable recommendations:

- Offer regular opportunities for professional development and skill enhancement.
- Provide clear paths for career progression and communicate these to employees.

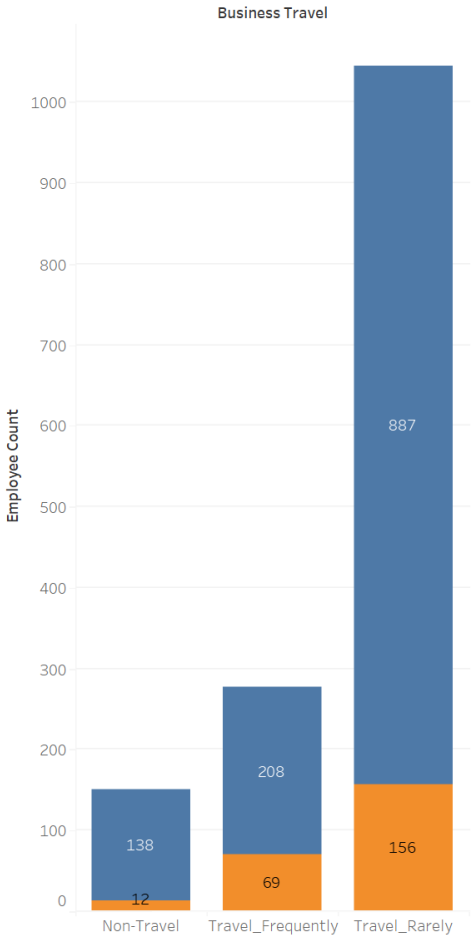


.4. Business Travel:

Frequent business travel (highest ratio of employees leaving to employees staying ~0.33) can disrupt work-life balance and cause stress. Employees who are required to travel extensively may become dissatisfied and consider other options.

Actionable recommendations:

- Review the absolute necessity of frequent travel and explore alternatives like virtual meetings when feasible.
- Consider offering flexible schedules or additional support for employees who travel frequently.
- Recognize the added burden of business travel and consider appropriate compensation or benefits.



.5. Stock Option Level:

Employees with a stock option level of 0 have the highest risk of attrition. Those with stock option levels of 1, 2, or 3 likely have a lower risk of leaving, with increasing retention potential as the stock option level grows. The ratio of employee staying to employee leaving is lower for stock option level 0 (~3) as compared to other stock level options (9.6,12.2 ,4.7 for stock level options 1,2,3 respectively), as can be seen from the visualisation below.

Actionable recommendations:

- Review stock option plans and their distribution across different employee levels.
- Consider expanding stock option programs to increase employees' financial stake in the company's success.