**Introduction: Examining Employee Turnover**

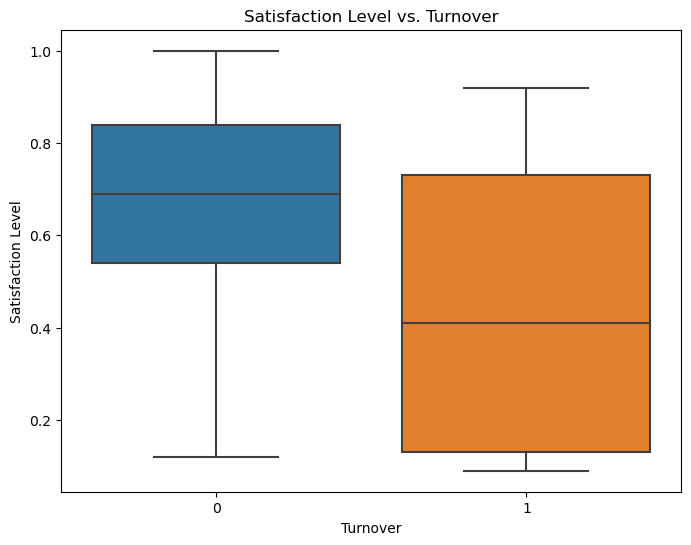
**By Matthew Latondresse**

**Problem Introduction:** The problem at hand revolves around predicting employee turnover, a critical concern for organizations. Employee turnover, or churn, refers to the departure of employees from a company and the need to replace them with new hires. This phenomenon can have significant implications for an organization's productivity, morale, and overall success.

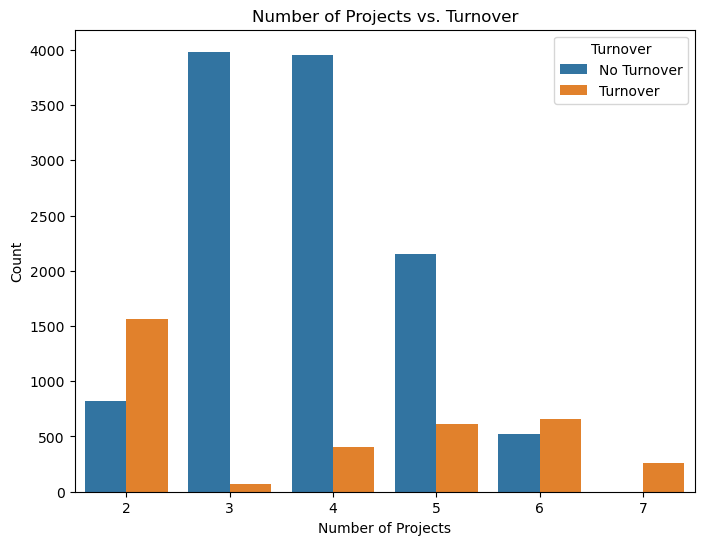
**Importance and Justification:** Employee turnover is a pressing issue for businesses across various industries. High turnover rates can lead to increased recruitment costs, disruptions in workflow, loss of institutional knowledge, and a negative impact on team dynamics. Moreover, the departure of skilled employees can hinder a company's ability to achieve its strategic goals. Addressing and predicting employee turnover is crucial for organizations aiming to maintain a stable, efficient, and innovative workforce.

**Pitch to Stakeholders:** In presenting this problem to stakeholders, it is essential to highlight the potential financial and operational consequences of high employee turnover. By framing it as a strategic challenge, stakeholders can understand that addressing this issue is not only a matter of employee satisfaction but a critical component of long-term business success. Demonstrating how predictive modeling can provide actionable insights and mitigate turnover risks will resonate with stakeholders, emphasizing the tangible benefits of addressing this problem.

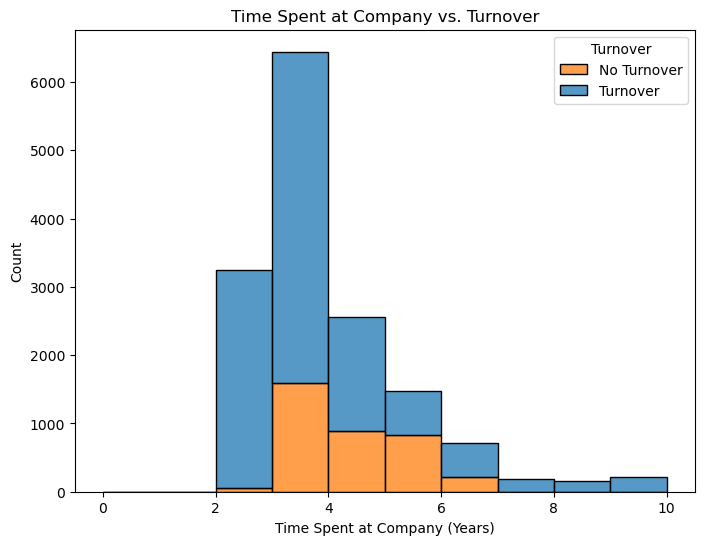
**Data Source:** This data set is from Kaggle. In a real-life scenario this data would be data obtained from an internal HR database, capturing information about employees' satisfaction levels, performance evaluations, number of projects, average monthly hours worked, time spent at the company, work accidents, promotions in the last five years, salary levels, workload, and department. The dataset offers a comprehensive view of numerous factors that might contribute to employee turnover. Ensuring data privacy and compliance with ethical standards, the source provides a valuable foundation for building a predictive model to identify potential turnover risks.  
  
**Problem Definition and Data Collection:** The problem of employee turnover was defined, emphasizing its significance for organizations. The dataset, containing information on various employee-related factors, was collected from an internal HR database. The goal was to predict which employees are more likely to leave the company. After identifying a few variables that could be potential reasons for employees leaving. Exploratory Data Analysis (EDA) was then done. This is a crucial step to understand the characteristics of the dataset. Key insights and visualizations were generated to explore the distribution of variables, identify patterns, and uncover potential relationships. Visualizations such as histograms, box plots, and correlation matrices were used to gain a comprehensive understanding of the dataset.



The first graph as seen above was created to examine the satisfaction of employees who had left versus those who were currently employed. This graph showed that the employees that are currently employed are expressing on average higher satisfaction than those that had left. Those who left expressed a much wider range of satisfaction levels.



The next graph that was created was to examine if the number of projects an employee was involved in was a potential factor in turnover. As seen in this graph employees seem most likely to stay when working on 3-5 projects.

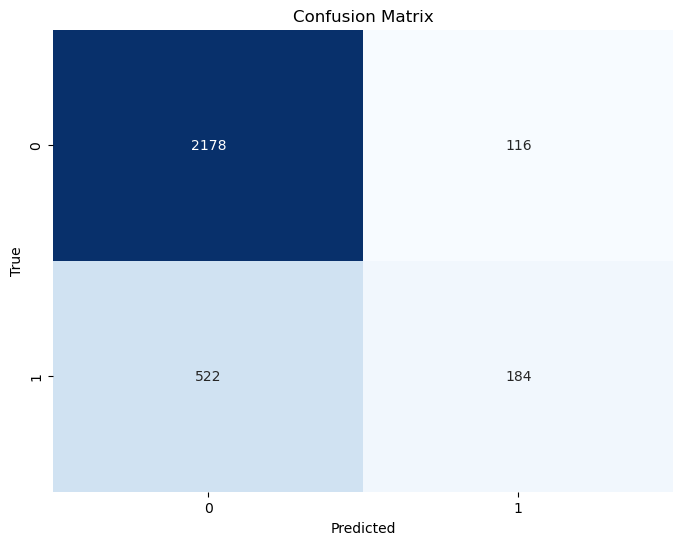


In the graph above time spent at the company was examined to see how long people stayed at the company before leaving. As shown on this graph an employee is more likely to leave the company after staying there for 3-5 years. Before that and after that point employees are more likely to stay.

**Data Preparation and Model Building** Data preparation included handling missing values, encoding categorical variables, and splitting the dataset into training and testing sets. The model selected for building was Logistic Regression, chosen for its interpretability and suitability for binary classification problems. The model was trained on the training set and evaluated on the testing set.

**Model Evaluation Summary:**

The Logistic Regression model, trained and evaluated on the employee turnover dataset, provides insights into its predictive performance. The following key metrics were used to assess the model:



* **Accuracy: 78.73%**
  + The overall accuracy of the model is approximately 78.73%. This metric indicates the proportion of correctly classified instances out of the total.
* **Precision: 61.33%**
  + Precision measures the accuracy of positive predictions made by the model. In this case, it suggests that when the model predicts an employee will leave, it is correct about 61.33% of the time.
* **Recall: 26.06%**
  + Recall, also known as sensitivity or true positive rate, signifies the proportion of actual positive instances that the model correctly identifies. A recall of 26.06% indicates that the model captures only about a quarter of the employees who left.
* **F1 Score: 36.58%**
  + The F1 score, a balance between precision and recall, provides a single metric to evaluate the model's performance. A score of 36.58% suggests a moderate trade-off between precision and recall.
* **ROC AUC Score: 60.50%**
  + The ROC AUC (Receiver Operating Characteristic - Area Under the Curve) score measures the model's ability to distinguish between classes. A score of 60.50% indicates a moderate discriminative capability.

**Insights:**

* The model exhibits a high accuracy, suggesting it performs well in overall predictions.
* The precision score indicates that when the model predicts an employee will leave, it tends to be correct.
* However, the recall score is comparatively low, indicating that the model may miss many employees who leave.
* The F1 score reflects the need for a balanced consideration of precision and recall, emphasizing the trade-off between false positives and false negatives.
* The ROC AUC score suggests a fair ability of the model to discriminate between employees who leave and those who stay.

**Conclusion:** While the model demonstrates reasonable performance, especially in terms of accuracy and precision, there is room for improvement in capturing true positives (recall). Further optimization, feature engineering, or considering alternative models might enhance the model's ability to identify employees at risk of turnover.

**Recommendations:**

* **Feature Importance Analysis:**
  + Conduct a thorough analysis of feature importance to identify which variables contribute most to predicting turnover. This can guide targeted interventions.
* **Model Tuning:**
  + Explore hyperparameter tuning or consider other machine learning algorithms to potentially improve model performance.
* **Employee Engagement Strategies:**
  + Based on the model insights, develop targeted employee engagement strategies for identified high-risk groups.

**Challenges and Opportunities:**

* **Challenges:** Addressing the imbalance between precision and recall achieving a more holistic model.
* **Opportunities:** Further exploring employee satisfaction dynamics and incorporating additional relevant features for a more comprehensive predictive model. As the model continues to evolve and improve it will better serve the company and its employees.