**Prediction of Employee Retention and Promotion Using Machine Learning**

Zachary Artman, Anand Fernandes, Jim McCarthy

University of San Diego

AAI-501-01 Introduction to Artificial Intelligence

Professor Andrew Van Benschoten, Ph.D

August 11, 2024

**Introduction**

In order to understand some of the causes of employee turnover and the likelihood that a person would be promoted, we ran a number of machine learning models that might be able to identify factors that influence an employee’s desire to stay with a company, and the company’s desire to keep the employee over the long term.

The dataset we used was obtained from contained nearly 15000 records and 10 different employee attributes, such as satisfaction level, average monthly hours, number of projects worked on, and salary.

Using different machine learning models, we planned to make predictions that could guide management decisions and improve employee satisfaction. To determine what features had the most predictive power, we analyzed the data using a combination of data visualization, correlation analysis, and predictive modeling.

**Data Preparation**

We began by processing the dataset and converting the categorical values into numerical ones to facilitate analysis. The machine learning models we used needed the feature variables to be numerical.

The three salary levels of low, medium, and high were mapped to numerical values of 0, 1, and 2 respectively. The department names were encoded as integers as shown (table 1)

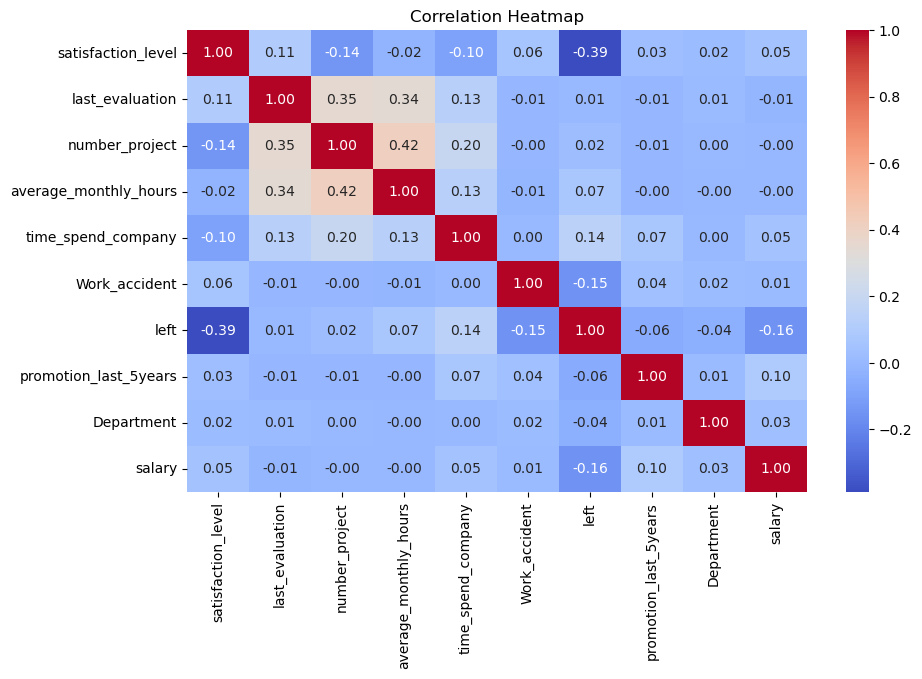
|  |  |
| --- | --- |
| Numerical Value | Categorical Value |
| 0 | Sales |
| 1 | Accounting |
| 2 | HR |
| 3 | Technical |
| 4 | Support |
| 5 | Management |
| 6 | IT |
| 7 | RandD |
| 8 | Product Management |
| 9 | Marketing |

Table 1. Map of Departments

**Data Analysis**

After cleaning the data, we performed a number of different tests to determine how different features related to one another, and which features had the most influence on whether an employee left the company or was promoted.

We started with a heatmap (figure #1) of the various features contained within the dataset. From this we found that satisfaction\_level had the most correlation of -0.39 with remaining with a company. The next highest correlations were salary level at -0.16 and whether the employee had a work place accident at -0.15. The highest correlation as predictor of whether the employee would leave the company was 0.14 for the time spent with the company, and 0.07 for the number of hours per month spent with the company.

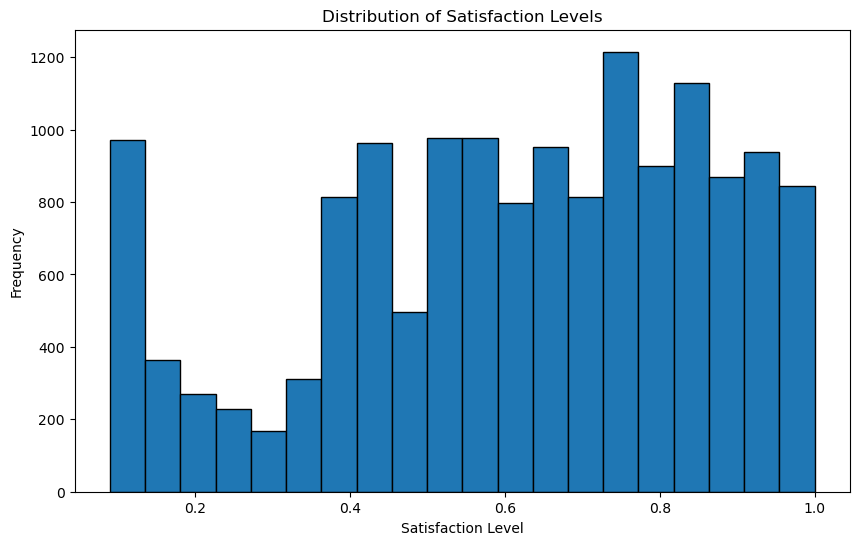


*Figure 1. Heatmap of Feature Correlation*

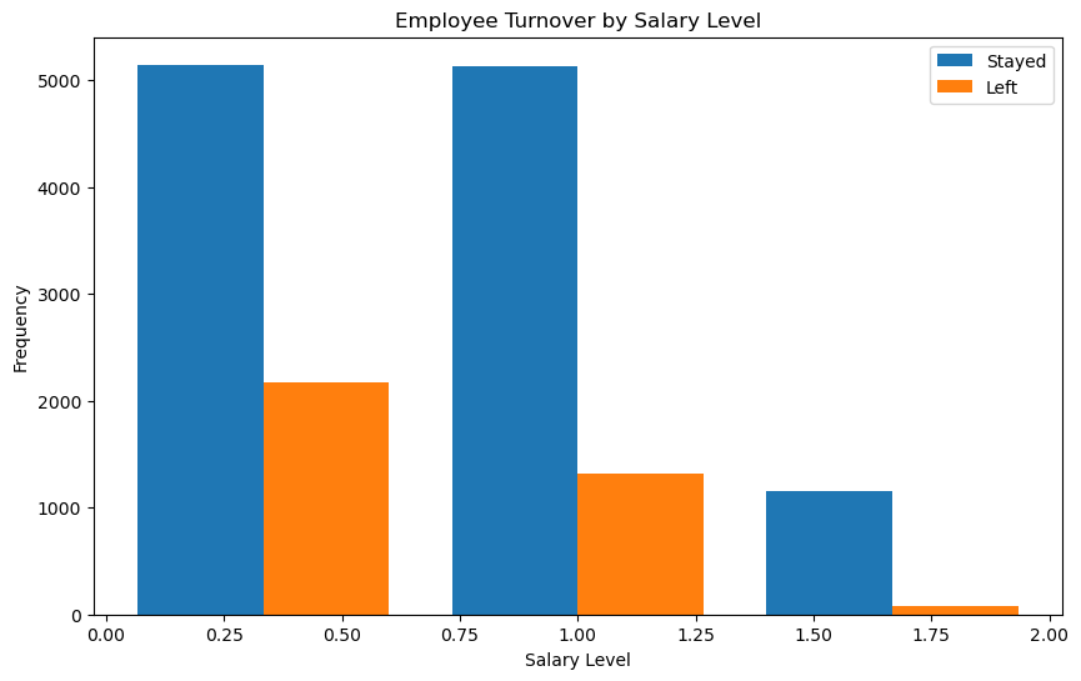
The features with the highest correlation were then checked individually to discover how they related to employee retention. A graph of employee turnover versus satisfaction level was run (figure 2), as well as a graph of just satisfaction level (figure 3), and a graph of retention versus salary level (figure 4). Other graphs run include promotion over the last 5 years, average monthly hours by department, average salary by department, and promotion versus retention. We also ran a plot of feature importance in predicting promotions.



*Figure 2. Turnover versus Satisfaction Level*



*Figure 3. Satisfaction Level*



*Figure 4. Turnover versus Satisfaction Level*

**Key Insights from Data Visualization**

The visualizations provided several important insights:

1) Each department has pretty much the same amount of monthly hours.

2) The number of people at each satisfaction level was fairly consistent, though there weren’t many people at levels between 0.2 and 0.4. There was a significant spike at 0.1 for those with a satisfaction level of less than 0.4. This would indicate that people either hate their job, are ambivalent, or like their job. Those that mildly dislike their job weren’t represented in the sample. Employees with lower satisfaction were more likely to leave the company, whereas those with mild dislike to high satisfaction were likely to remain at the company.

3) The data showed that on average only 2% of the people were promoted in the last 5 years. Review of other data outside this sample indicates that this value is rather low, with common promotion rates more likely around 6% (Society for Human Resource Management, 2016).

4) There is a significant relationship between satisfaction level and whether someone would leave or stay with the company. The higher the satisfaction level, the more likely a person would remain with the company.

5) The average monthly hours and number of projects weren’t good predictors of whether a person would leave the company, but the average number of hours worked was a good predictor of predicting promotion.

Departments like management had higher average salaries, while average monthly hours remained consistent across departments. This suggests that factors other than workload, such as role or responsibility, contribute more significantly to salary variations.

**Predictive Modeling**

Several predictive models were developed to address the key questions posed. An analysis of different models showed that the RandomForest classifier was best suited to predict employee turnover, employee promotions, and employee salary.

A secondary analysis was made regarding the prediction of a work place accident. The RandomForest model was good at identifying non-accidents, while SVC was better at detecting actual accidents, even though it had many false positives. Stacking these models show that the stacking model kept the high accuracy in predicting non-accidents, similar to the RandomForest, while slightly improving the recall for accidents compared to the RandomForest alone. This means that the stacking model is a bit better at catching actual accidents without sacrificing much accuracy in other areas.

There were a few features that had an important effect on predicting employee promotions (figure 6), with the average monthly hours being the greatest predictor. When looking at the results from the model predicting promotions, we see that the precision for predicting promotions is perfect at 1.00, so all predicted promotions were correct. However, the recall for promotions is lower at 0.56, indicating that the model missed more than half of the actual promotions. In other words, if the model predicts that someone will get a promotion, they almost certainly will. But the model often fails to identify all employees who will get promoted, so it may incorrectly predict that some individuals will not get a promotion even though they will.

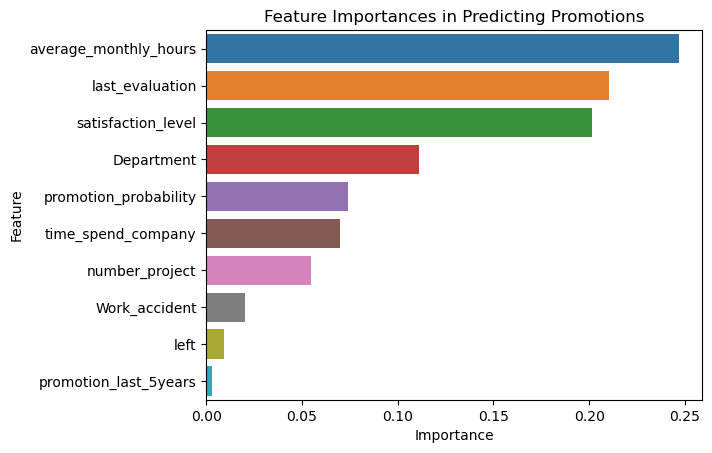
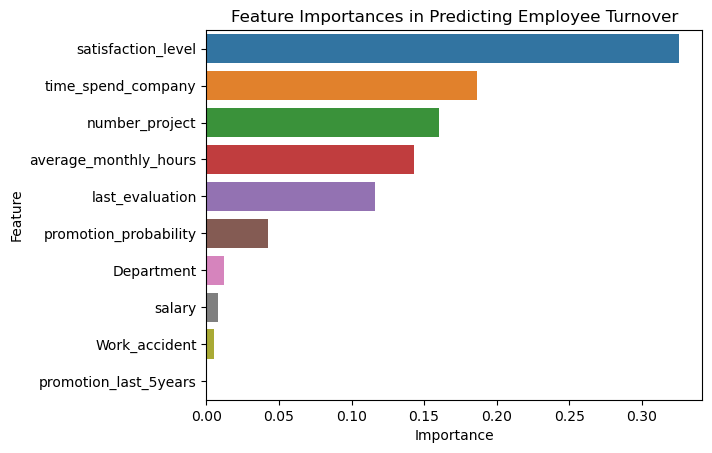


Figure 6. Feature Importance in Predicting Promotions

When predicting employee turnover, the employee satisfaction level is the feature with the greatest weight in predicting whether the employee will stay with or leave the company (figure 7).

The RandomForest classifier performs very well when predicting employee turnover, with high precision and recall for predicting those who stayed and those who left. The recall for predicting those who left (Class 1) is slightly lower, meaning the model misses a small number of actual leavers. The model's accuracy suggested that it could be a useful tool for predicting employee attrition and helping management take preemptive actions.



. Figure 7. Feature Importance in Predicting Employee Turnover

Lastly, a model was developed to predict employee salaries. The model performed reasonably well in predicting low and medium salaries but struggled significantly with high salary predictions, indicating a bias towards the majority salary levels in the dataset. This bias suggests that the model may need further refinement or additional data to accurately predict higher salary levels.

The features most important to the model (figure 8) indicate that it closely matches the features that are important to getting a promotion. Average monthly hours, whether the person has had an evaluation, and employee satisfaction level match the top three features of importance for employees that get a promotion.

The salary prediction model performed reasonably well in predicting low salaries (class 0), with a precision of 0.77 and recall of 0.75, meaning it correctly identifies most employees with low salaries while maintaining relatively few false positives. For medium salaries (class 1), the precision is slightly lower at 0.68, but the recall is the same at 0.75, indicating that while the model captures most employees with medium salaries, it has a higher rate of false positives, incorrectly predicting medium salary for some employees. However, the model struggles significantly with predicting high salaries (class 2), where the recall drops to 0.38, meaning it misses a large proportion of actual high-salary employees, despite having a decent precision of 0.74. While the model can often correctly identify high salaries when it predicts them, it oftentimes fails to recognize many employees who actually have high salaries.

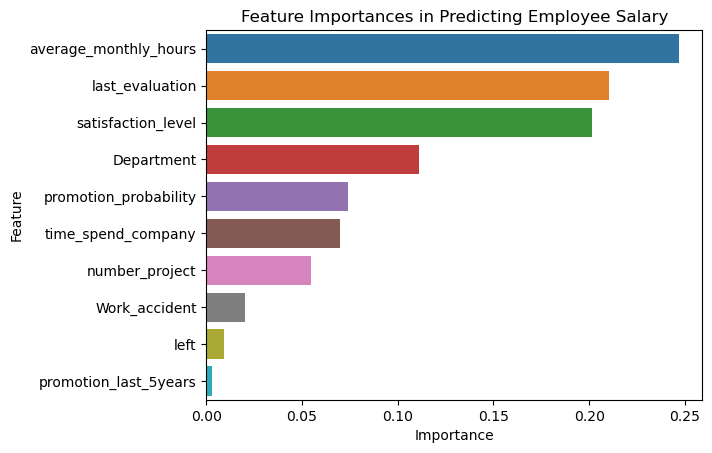


Figure 8. Feature Importance in Predicting Employee Salary

**Discussion and Implications**

The findings from this analysis have several important implications:

1. Improving Promotion Rates: The low promotion rate within the dataset may contribute to lower satisfaction and higher turnover. If the data represents a single company, then the company should consider plans for improving identification and promotion of its employees.
2. Addressing Employee Satisfaction: The strong correlation between low satisfaction and turnover highlights a likely need for targeted interventions to improve employee morale. The large number of individuals have a satisfaction level of less than 80%. An average employee satisfaction of 80% is the benchmark for companies that are considered the “best places to work”.
3. Salary Disparities: Though it is common for management to have higher salaries, it is important to ensure that the salary disparity is not too large. A more transparent and equitable salary structure could help address this issue.

**Conclusion**

We successfully applied machine learning techniques to analyze and predict key employee outcomes. While the models showed some promise, particularly in predicting turnover and promotions, challenges remain, particularly in predicting salary levels. Future datasets could provide better feature richness to refine the models, and additional data from other datasets could be incorporated to compare and contrast against. Exploring other machine learning algorithms to improve prediction accuracy would also be useful.

**References**

Society for Human Resource Management. (2016). Customized human capital benchmarking

report. <https://shrm.org/benchmarks>

**Appendix A**

For access to the code and data used in this study, visit the following GitHub repository: <https://github.com/zacharyartman/aai501-group2>.