Employee Turnover Analysis and Prediction

Introduction

The goal of this analysis is to predict employee turnover at Portobello Tech and understand the key factors contributing to employees leaving the company. We will be using a dataset (HR_comma_sep.csv) to evaluate patterns related to employee satisfaction, performance evaluations, and turnover. The analysis involves performing feature exploration, correlation analysis, K-Means clustering, and Logistic Regression for predicting employee turnover.

2.1 Correlation Analysis

In this step, we examined the correlation between different features in the dataset to better understand the relationships between them. Using a heatmap, we observed the following key correlations:

- **Satisfaction level** has a negative correlation with employee turnover (left), meaning that lower satisfaction is linked with higher turnover.

- **Last evaluation** shows a moderate negative correlation with turnover as well. This suggests that employees who have lower evaluations are more likely to leave.

- **Work accidents**, **promotion last 5 years**, and other features had weaker correlations, indicating they are less relevant to predicting turnover.

The correlation matrix showed that the most significant predictor of turnover is **satisfaction level**, followed by **last evaluation**.

3.1 Logistic Regression

To build a model predicting whether an employee will leave the company, we used Logistic Regression. The features used for training included satisfaction_level, last_evaluation, number_project, average_montly_hours, time_spend_company, Work_accident, promotion_last_5years, and sales.

The model was trained on a training set and evaluated using cross-validation, with the following results:

- **Average accuracy**: 0.74

- **Cross-validation scores**: The model achieved reasonable performance in predicting turnover, with an average accuracy of 74%. However, this accuracy might be improved

by fine-tuning the model or using additional features.

The coefficient analysis showed that employees with lower satisfaction and poor evaluations were more likely to leave, supporting the findings from the correlation analysis.

3.2 K-Means Clustering

For further insights into turnover, we performed K-Means clustering on employees who left the company, using satisfaction_level and last_evaluation as features. The clustering revealed three distinct groups of employees:

1. **Cluster 0 – Moderate Satisfaction, Low Evaluation**: Employees in this cluster had moderate satisfaction (~0.4) and low performance evaluations. They are likely to have been underperforming and may have left due to dissatisfaction with their job roles, performance, or lack of advancement opportunities.

2. **Cluster 1 – High Satisfaction, Low Evaluation**: Employees in this group had low satisfaction levels but performed well according to their evaluations. This suggests that some high-performing employees might leave due to dissatisfaction with their role, such as lack of recognition, unchallenging tasks, or poor work-life balance despite good performance.

3. **Cluster 2 – High Satisfaction, High Evaluation**: This group, surprisingly, consists of employees who were both satisfied and high performers. Even though these employees had high satisfaction and good evaluations, they left the company, indicating that external factors such as better career opportunities or higher salaries might have influenced their decision to leave.

The findings from the clustering analysis suggest that employee turnover occurs due to a variety of factors, and not just dissatisfaction. Employees with high evaluations and satisfaction may leave for reasons unrelated to their current roles, while underperforming employees with moderate satisfaction are also likely to exit due to dissatisfaction.

3.3 Logistic Regression with Cross-Validation

The Logistic Regression model was then validated using cross-validation. The cross-validation results confirmed that the model's performance was consistent across different folds, with an average accuracy of 74%. This reinforced the conclusion that satisfaction level and last evaluation are key predictors of turnover.

Conclusion

The analysis of employee turnover at Portobello Tech revealed important insights into the factors influencing turnover. By analyzing correlations, clustering, and using Logistic Regression, we identified that employees with low satisfaction and low evaluations are more likely to leave. However, we also found that employees with high evaluations and satisfaction may leave due to external factors like better job opportunities. The Logistic Regression model achieved an accuracy of 74%, indicating reasonable predictive power.

The findings from this analysis can help Portobello Tech better understand employee turnover and create targeted strategies to improve employee retention. Employees with lower satisfaction or evaluations may benefit from career development opportunities, while high-performing employees could be better engaged by offering them career advancement and recognition.