Predicting Employee Turnover: A Machine Learning Approach

Joanna John

North Carolina State University

DSC412: Exploring Machine Learning

Trenton Embry

November 26, 2024

Abstract

Employee turnover is a critical challenge for organizations aiming to retain talent and reduce associated costs. This project investigates key drivers of employee turnover using the Employee Turnover Analytics Dataset. By analyzing employee characteristics and organizational factors, we develop predictive models to identify employees at risk of leaving. Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN) models were evaluated, with Decision Tree achieving the highest accuracy (97.92%). The results provide actionable insights for employers to improve employee retention.

Background

Employee turnover imposes significant financial and operational burdens on organizations, including recruitment costs and decreased productivity. Understanding the factors influencing turnover can help HR professionals and business leaders implement data-driven strategies to retain valuable employees. This project leverages machine learning techniques to analyze employee data, identify key factors linked to turnover, and predict employee exit risks.

Data Analysis

The dataset was sourced from Kaggle and comprised both categorical and numerical features. Data exploration involved statistical summaries and visualizations, aiming to uncover patterns associated with turnover. Key features examined included satisfaction level, evaluation scores, project counts, and salary categories. Some important observations I made are:

- Satisfaction Level: Low satisfaction levels strongly correlated with higher turnover rates.
- Last Evaluation Scores: Employees with extremely low or high evaluations were more likely to leave.

- Number of Projects: Employees engaged in too few or too many projects had higher turnover rates.
- Average Monthly Hours: Employees working longer hours exhibited higher turnover, highlighting potential burnout.
- Years at the Company: Turnover peaked at around five years of tenure.
- Salary Levels: Lower salaries were strongly linked to higher turnover rates.

These observations guided feature engineering, such as creating bins for continuous variables and selecting relevant predictors. Features with strong correlations to turnover were prioritized for modeling.

Data Augmentation

Cleaning and Preparation

- Missing Values: No missing values were identified, simplifying preprocessing.
- Column Renaming: "Sales" was renamed to "Department," and "Average Monthly Hours" was converted to "Average Weekly Hours."
- Feature Selection: Non-predictive features like department were removed, and categorical variables were encoded numerically.

Feature Engineering

- Binning: Continuous variables like evaluation scores and project counts were categorized into ranges for interpretability.
- Normalization: Numerical features were scaled for compatibility with modeling algorithms.

Model Selection

Three models were selected based on their suitability for classification tasks:

- Logistic Regression: To interpret feature coefficients and their impact on turnover probability.
- K-Nearest Neighbors (KNN): For its simplicity and effectiveness in non-linear relationships.
- Decision Tree: For its interpretability and ability to capture complex patterns.

The Decision Tree model was favored for its interpretability and high accuracy. Logistic Regression provided valuable insights into feature importance, while KNN served as a baseline model.

Training Methodology

The dataset was split into 80% training and 20% testing subsets to ensure robust model evaluation. When it came down to tuning hyperparameters, I had to do it for the following 2 models:

- KNN: The optimal number of neighbors was determined through testing, with k=3 yielding the best results.
- Decision Tree: The maximum depth was adjusted to balance complexity and overfitting.

Cross-validation was employed to evaluate model performance and prevent overfitting.

Results

Model	Accuracy (%)
Logistic Regression	79.34
KNN	85.23
Decision Tree	97.92

The Decision Tree model emerged as the best-performing model, providing clear interpretability and actionable insights.

The Decision Tree revealed the most critical factor: an abnormal number of projects significantly increases turnover risk. Logistic regression coefficients highlighted the protective effect of higher satisfaction levels and salaries, as well as the harmful effect of working on an excessive number of projects (both too few or too many) and having a stellar/poor evaluation.

Future Work

To enhance the project, several improvements could be made. Data enrichment through incorporating additional variables, such as detailed performance metrics or employee survey data, could provide a more comprehensive analysis of turnover predictors. Model optimization could be achieved by experimenting with ensemble methods like Random Forests or Gradient Boosting to further improve prediction accuracy. Collaborating with HR professionals to validate model predictions and refine feature selection based on domain expertise would also strengthen the model's relevance and usability. For future directions, developing an interactive tool for HR departments to visualize turnover risks would make the insights more accessible and actionable. Additionally, expanding the scope of analysis to include industry-specific turnover trends could provide broader and more contextual insights for various sectors.

Stakeholder Acknowledgments

The insights derived from this project benefit multiple stakeholders. HR departments can leverage the findings to develop targeted retention strategies, while business leaders can use the predictions to reduce turnover costs and enhance workforce stability. Recruiting teams can anticipate and address potential talent shortages more effectively, and employees may benefit from improved workplace policies informed by data-driven decisions. However, it is crucial to use the model's predictions responsibly to avoid biases or unintended consequences, such as unfairly stigmatizing certain employees.

Conclusion

This project demonstrates how machine learning can analyze employee data to predict turnover effectively. By leveraging insights into critical factors like project assignments and satisfaction levels, organizations can proactively address turnover risks. The Decision Tree model's high accuracy highlights its potential as a decision-making tool for HR departments.

References

[1] A. Hedau, “Employee turnover analytics dataset,” Kaggle, https://www.kaggle.com/datasets/akshayhedau/employee-turnover-analytics-dataset/data (accessed Nov. 26, 2024).

[2] “Cite a website in IEEE,” Citation Machine, a Chegg service, https://www.citationmachine.net/ieee/cite-a-website (accessed Nov. 26, 2024).