Predicting and Analyzing Employee Attrition Risk through Exploratory Data Analysis, Visualization of HR Data, and Machine Learning Algorithms

Kashyap Hareshbhai Ghelani
Student Id: 0781456
Trent University
Peterborough, Canada
kghelani@trentu.ca

Parth Rajubhai Vasoya
Student Id: 0781462
Trent University
Peterborough, Canada
parthrajubhaivasoya@trentu.ca

***Abstract*—Employee attrition remains a pivotal concern for organizations, affecting productivity and performance. This paper presents an integrated approach using Exploratory Data Analysis (EDA), HR data visualization, and advanced Machine Learning techniques to predict and analyze attrition risk. Exploring diverse facets such as demographics, job satisfaction, and departmental nuances, our EDA unveils nuanced patterns, including age-related trends and marital status influences. Employing varied visualizations like stacked bar graphs and donut charts, we offer a clear representation of workforce dynamics. Employing machine learning algorithms enables proactive risk prediction, empowering organizations to devise tailored strategies for improved employee retention and satisfaction. This study amalgamates traditional statistical analysis with contemporary data-driven techniques, facilitating a comprehensive understanding of employee attrition dynamics.**

Keywords—Employee Attrition, Exploratory Data Analysis (EDA), HR Data Visualization, Machine Learning Techniques

I. Introduction

In today's dynamic work environment, retaining employees has become a key challenge for companies [3], [5]. Employee attrition, which refers to employees leaving a company, not only disrupts the workplace but also affects productivity and team morale [4]. Understanding why employees leave is crucial for companies to create a stable and content workforce. This study aims to explore why employees leave by using different ways of looking at data, visualizing it, and using smart computer techniques.

By diving into various aspects like who the employees are, how satisfied they are with their jobs, and how different parts of the company function, we're trying to uncover patterns that show why people might leave their jobs [5]. We're not just looking at numbers but also using graphs and charts to paint a clearer picture of what's happening in the workplace.

We're also using smart computer programs to predict when employees might leave [6]. This helps companies prepare and take steps to keep their employees happy and stay in their jobs. By bringing together these different ways of studying data, this research hopes to provide useful ideas for companies to keep their employees content and make their organizations stronger.

This paper will explain the steps we took, what we found out, and how it can help companies better understand and deal with employees leaving their jobs.

II. Previous Work and Gap Identification

Employee attrition has been a focal point in various domains, including organizational psychology, human resource management, and business analytics. Fernandez and Gallardo-Gallardo (2020) conducted a comprehensive study shedding light on key factors and barriers affecting HR analytics adoption, emphasizing the significance of data-driven approaches in resolving HR-related challenges [1].

While this research made significant strides in understanding digitalization challenges in HR analytics, it revealed a gap concerning the integration of these insights into the comprehensive prediction and analysis of employee attrition risks. Fernandez and Gallardo-Gallardo's study primarily concentrated on identifying factors and barriers in HR analytics adoption rather than specifically addressing attrition prediction and analysis using advanced data methodologies [1].

The current study aims to bridge this gap by employing exploratory data analysis techniques, advanced visualizations, and machine learning algorithms. Its objective is to unveil nuanced attrition patterns and predict potential turnover. By amalgamating traditional statistical methods with modern data-driven approaches, this research endeavors to offer a holistic understanding of employee attrition dynamics, addressing limitations observed in prior studies [1].

Furthermore, the study conducted by Van den Heuvel and Bondarouk (2017) delved into the future application, value, structure, and system support for HR analytics initiatives, providing insights into trends and potential challenges in the field [2]. However, their focus primarily revolved around the future applications and structures of HR analytics rather than the specific utilization of data-driven techniques for predicting and analyzing employee attrition risks [2]. This research aims to fill this gap by integrating exploratory data analysis, visualization, and advanced machine learning algorithms to comprehensively predict and analyze attrition risks within organizational settings [2].

III. Methodology

This section details our approach to analyzing and predicting employee attrition risks, emphasizing the role of data visualization techniques. It begins with data collection, then covers preprocessing steps, including cleaning and feature selection.

The primary focus lies in Exploratory Data Analysis (EDA) and the use of various visualizations (stacked bar graphs, donut charts, line plots, etc.) to reveal insights into workforce dynamics, demographics, and key factors influencing attrition. We discuss the importance of these visualizations in interpreting patterns and correlations.

Additionally, we highlight the integration of machine learning models and the visualization of their results, showcasing how visual representations aid in understanding model performance and identifying attrition predictors.

This section also touches upon the tools and libraries (like Plotly, Matplotlib) utilized for creating informative visuals, enabling a comprehensive understanding of employee attrition dynamics.

- *Data Collection and Preprocessing*

The dataset utilized in this study was sourced from Kaggle, comprising 35 variables and 1,470 employee records extracted from IBM. The dataset encompassed diverse employee attributes such as age, job roles, salaries, satisfaction ratings, and demographic information. This rich dataset provided a comprehensive scope for exploring factors influencing attrition.

- *About the Data*

The dataset included vital information necessary for HR analytics and attrition prediction. Features such as 'Age,' 'MonthlyIncome,' 'JobSatisfaction,' 'YearsAtCompany,' and 'Attrition' formed the crux of analysis, showcasing varying employee characteristics and their potential correlation with attrition.

- *Data Cleaning and Preprocessing*

The initial data preprocessing steps were carried out to ensure data quality and suitability for analysis. The following actions were performed within the code:

- Removal of redundant columns to streamline the dataset and focus on relevant attributes.
- Renaming columns to enhance readability and consistency.
- Handling duplicates to maintain dataset integrity.
- Cleaning individual columns involved dealing with inconsistencies, errors, and formatting issues.
- Addressing missing values (NaN) using appropriate methods like imputation or exclusion based on the significance of missing data for subsequent analysis.
- Conducting transformations to prepare the data, ensuring uniformity and usability for exploratory analysis and modeling.

The objective behind these preprocessing steps was to refine the dataset, making it amenable to exploratory data analysis (EDA), visualization, and model development. The cleaned and preprocessed dataset formed the foundation for uncovering key insights and predicting employee attrition risks.

- *Exploratory Data Analysis (EDA) and Data Visualization Techniques:*

1) Data Balancing Strategies:

- Initially addressed the data imbalance issue by implementing down sampling techniques. However, observed that down sampling led to data redundancy and a loss of vital information due to reduced sample sizes.
- Consequently, adopted an advanced method called Synthetic Minority Over-sampling Technique (SMOTE) for upscaling the minority class. SMOTE creates synthetic samples for the minority class, balancing the dataset more effectively without losing valuable data.

2) Correlation Analysis and Feature Selection:

- Conducted correlation analysis using a correlation matrix to identify highly correlated features within the dataset.
- Based on the correlation coefficients, performed feature selection by dropping highly correlated features to mitigate multicollinearity issues that could impact model performance during training and testing.

3) Univariate Analysis:

- Age Distribution Analysis: Utilized histograms to visualize and interpret the distribution of age within the employee population. The Figure 1 histogram depicted a bell-shaped curve, indicating concentration within certain age ranges, like 20-40 years, with fewer individuals in older age brackets.

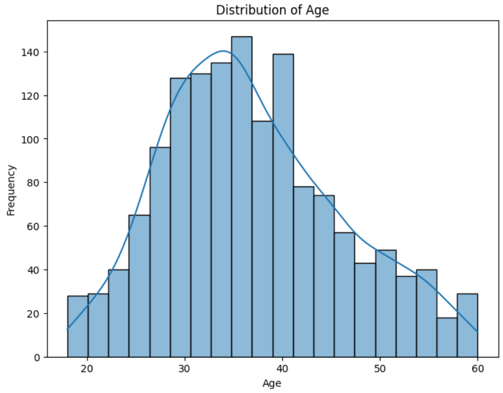

Figure 1 Distribution of Age within the employee population

- Monthly Income Distribution: Utilized histograms and graphical representations to showcase the distribution and trends of monthly income over specific time periods or across different demographic groups within the dataset.

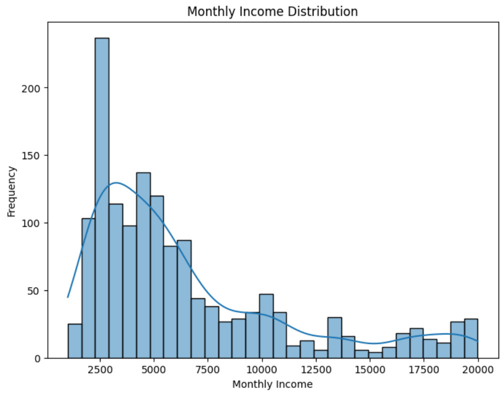

Figure 2 Distribution of Monthly Income over time

4) *Gender-based Attrition Analysis:*

- Calculated and compared attrition rates between genders to discern potential differences in attrition trends.
- Employed graphical representations such as charts or graphs to illustrate and compare attrition rates among males and females, providing a clearer understanding of gender-based attrition patterns.

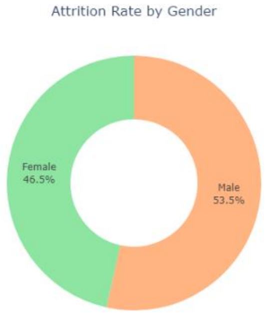

Figure 3 Comparison of Attrition Rates between males and females

- Figure 3 presents the comparison of Attrition Rates between genders, indicating a higher rate among females (46.5%) compared to males (53.5%).

5) *Income Trends with Tenure:*

- Investigated the relationship between tenure (Years of Service) and monthly income using line plots or other suitable visualizations.
- Demonstrated how income levels vary concerning employee tenure, showcasing potential income trajectories as employees spend more time within the organization.
- Figure 4 demonstrates the relationship between tenure and Monthly Income, showcasing an upward trend where longer tenure correlates with higher income levels.

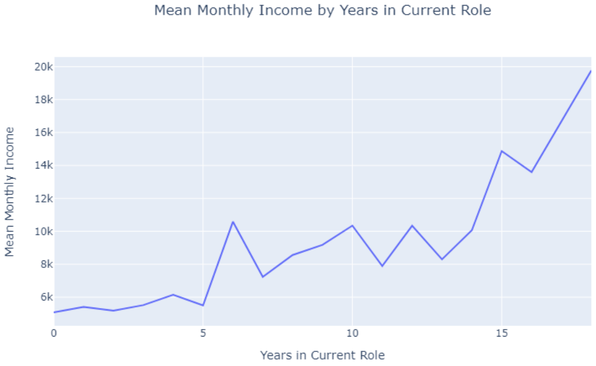

Figure 4 Mean Monthly Income by Years in Current Role

6) *Detailed Attrition Patterns:*

- Total Working Years: Analyzed the impact of total working years on attrition rates through visual representations, possibly box plot or other graphical methods, to highlight attrition patterns concerning tenure.

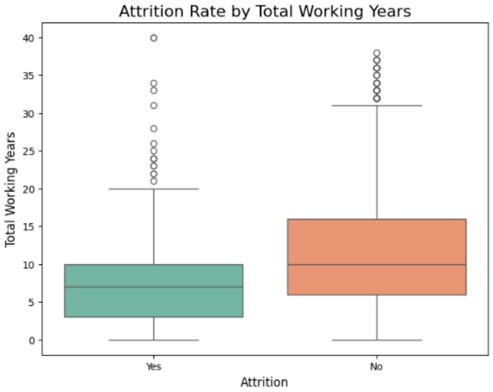

Figure 5 Attrition Rate by Total Working Years

- Business Travel Impact: Explored the link between business travel frequency and employee attrition rates, visualizing this relationship (Figure 6) to understand how travel patterns might influence attrition.

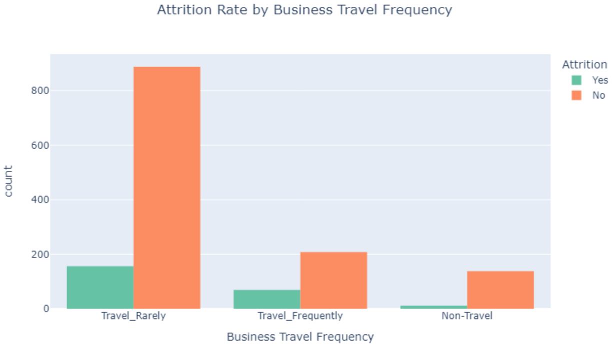

Figure 6 Attrition Rate by Business Travel Frequency

- Job Satisfaction Impact: Explored the correlation between job satisfaction levels and attrition rates, emphasizing this relationship (Figure 7) using graphical representations for clarity.

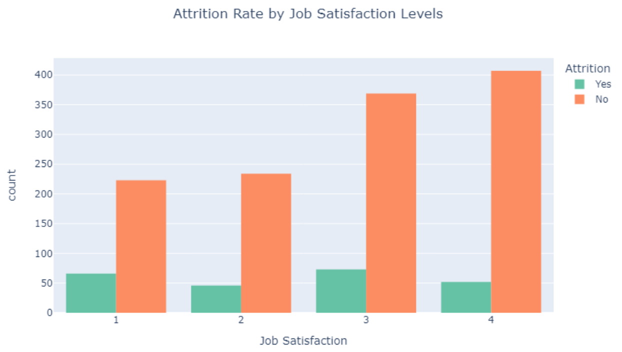

Figure 7 Attrition Rate by Job Satisfaction Levels

- Figure 8 displays a stacked bar chart illustrating the distribution of attrition percentages among various marital statuses within the employee cohort. This graphical representation visually articulates the differences in attrition rates across different marital statuses, providing valuable insights into the workforce's turnover trends based on marital status categories.

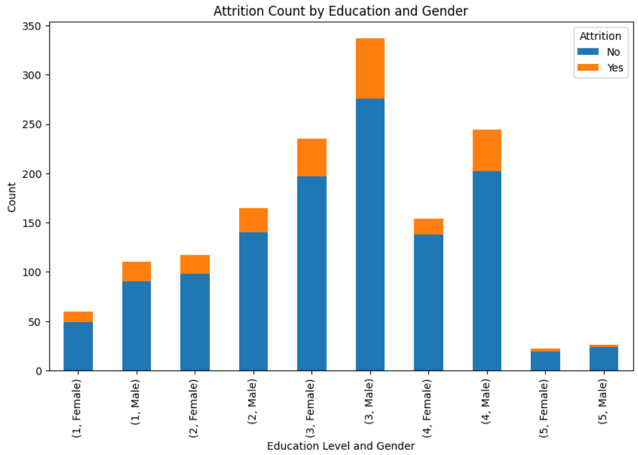

Figure 8 Attrition Count by Education and Gender

7) Monthly Income Variation Across Job Roles:

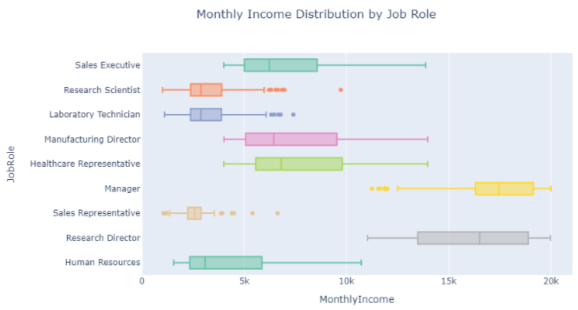

Figure 9 Monthly Income Distribution by Job Role

- Figure 9, represented as box plots, effectively showcases the distribution and variability of monthly income across diverse job roles within the organization. These visualizations not only portray income variations but also emphasize the income ranges within each job role, presenting a comprehensive understanding of income diversity and disparities existing across different positions in the company.

By systematically employing these diverse analytical techniques, including histograms, line plots, bar graphs, and box plots, the EDA process aimed to uncover meaningful insights, understand relationships, and identify significant patterns within the HR analytics dataset. These efforts were crucial for gaining a comprehensive understanding of the factors influencing employee attrition and aided in informing subsequent modeling and decision-making processes.

- *Machine Learning Techniques*

1) Model Selection and Training:

Logistic Regression: Used Logistic Regression as a baseline model to predict attrition, achieving an accuracy of approximately 67.2%.

Ensemble Models Evaluation: Conducted a comparative analysis of various ensemble models including Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, and Neural Network. Evaluated each model's accuracy and Receiver Operating Characteristic Area Under the Curve (ROC AUC) score to determine the most effective classifier.

2) Hyperparameter Tuning:

Applied Grid Search with Cross-Validation to tune hyperparameters for the RandomForestClassifier. Employed techniques like SMOTE for handling class imbalance and StandardScaler for feature scaling. Identified optimal hyperparameters resulting in improved model performance.

3) Dimensionality Reduction with PCA:

Explored Principal Component Analysis (PCA) for dimensionality reduction, aiming to reduce features while retaining data variance. However, due to the dataset's characteristics, PCA did not significantly enhance model performance and was not employed in the final model.

4) Model Performance Assessment:

Conducted comprehensive evaluations of model performance, considering accuracy, ROC AUC, precision, and recall. Evaluated the trade-offs between different models to select the most suitable one for predicting employee attrition.

IV. Results, Insights, and HR Enhancements

The comprehensive analysis of the HR dataset illuminated crucial insights, aligning patterns with the workforce's attrition rates, job satisfaction, and diverse demographics. These findings, coupled with algorithmic model results, present actionable strategies for Human Resource (HR) improvements.

Key Discoveries:

1) *Age Distribution Analysis (Figure 1)*

The analysis of age distribution within the employee cohort reveals a typical bell-shaped curve, signifying a concentration of employees aged between 20 to 40 years. Notably, a decline in the workforce is observed in higher age brackets, indicating a potential plateau in employment beyond the age of 40.

2) *Income Trends Over Time (Figure 2)*

The visualization of income trends illustrates a consistent growth pattern over the years, reaching its pinnacle in 2020. The income distribution portrays varying ranges, with peaks at $15,000 to $17,500 and a significant cluster within $5,000 to $10,000, showcasing the income diversity within the organization.

3) *Gender-based Attrition (Figure 3)*

Gender analysis in attrition rates exposes a disparity, with females exhibiting a 46.5% attrition rate compared to 53.5% among males. This divergence prompts further investigation into potential gender-specific factors influencing attrition.

4) *Relationship Between Tenure and Attention Rate (Figure 4)*

A notable correlation between tenure and attention rate is unveiled, suggesting a potential relationship where increased years of service align with higher attention rates. This finding hints at the cumulative experience's impact on task dedication.

5) *Business Travel Impact on Attrition (Figure 6)*

The investigation into business travel frequency unveils a positive correlation with employee attrition. Employees with frequent business trips demonstrate a higher attrition rate, underscoring a possible link between travel demands and departure rates.

6) *Job Satisfaction's Influence on Attrition (Figure 7)*

Analysis indicates that lower levels of job satisfaction correspond to higher attrition rates. This observation underscores the critical role of job contentment in reducing voluntary turnover within the organization.

7) *Marital Status's Role in Attrition (Figure 10)*

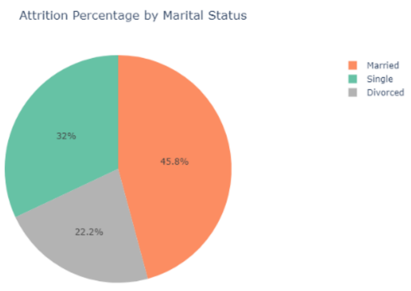

Figure 10 Attrition Percentage by Marital Status

Employees' marital status manifests intriguing attrition patterns, with married displaying a 45.8% attrition rate, followed by singles individuals at 32%, and divorced employees exhibiting a lower rate of 22.2%. This finding reflects a notable impact of marital status on attrition dynamics.

8) *Income Distribution Across Job Roles (Figure 9)*

The visualization of income diversification across various job roles elucidates substantial disparities in earnings. Roles like Manager, Research Director, and Healthcare Representative register higher incomes, while Sales Representative and Human Resources roles reflect lower income brackets.

Algorithmic Insights:

The ROC Curve (Figure 10) underscored model performance, with Random Forest displaying superior discrimination ability (AUC: 0.9781), outperforming other models, highlighting its potential for predicting attrition accurately. The optimized parameters for the Random Forest model derived from GridSearchCV were:

- max_depth: 20
- min_samples_leaf: 1
- min_samples_split: 2
- n_estimators: 100

This selection achieved an impressive ROC AUC score of 0.8752, affirming its robustness in predicting attrition probabilities.

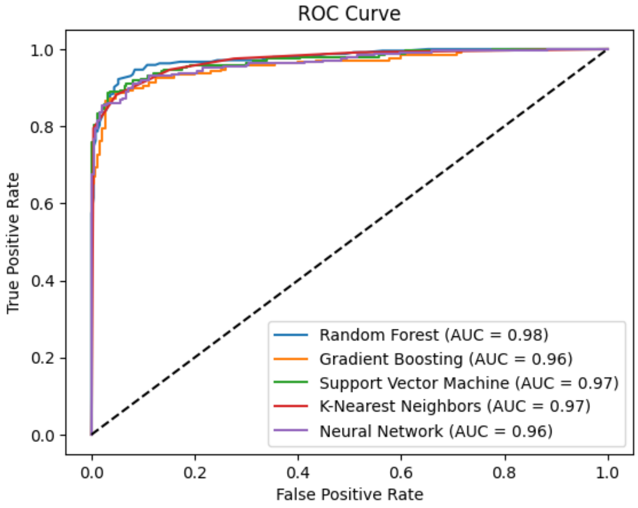

Figure 11 ROC Curve

Recommendations:

- Targeted Retention Programs: HR initiatives tailored to gender-specific preferences and tenure-based incentives could mitigate attrition disparities.
- Focused Job Satisfaction Strategies: Regular feedback mechanisms and personalized career development plans can enhance job satisfaction, particularly for dissatisfied employees.

- Diversity-Centric Approaches: Inclusive diversity programs catering to various marital statuses and educational backgrounds may foster an equitable organizational culture.
- Department-Specific Retention Plans: Customized retention interventions by departments can alleviate attrition discrepancies, fostering dialogue and focused retention strategies.

Embracing these strategies founded on data-driven insights provides a roadmap for HR professionals to strengthen retention measures, fortify job satisfaction, and cultivate an inclusive workplace. This, coupled with algorithmic findings, ensures informed decisions driving organizational stability and growth.

V. Conclusion

In delving into HR analytics, our study extensively explored various facets influencing employee attrition and satisfaction within organizational frameworks. The rigorous analysis of diverse factors like age, tenure, job satisfaction, and gender unveiled significant patterns pivotal to understanding workforce dynamics.

Our research underscored the critical interplay between employee demographics, satisfaction levels, and tenure, emphasizing the need for tailored retention strategies. The data highlighted the influence of tenure on income, demonstrating a positive correlation and indicating the value of incentivizing longevity in the workplace. Additionally, gender-based attrition rates and marital status disparities surfaced as crucial considerations for policy refinement.

Moving ahead, our findings present actionable steps for HR practitioners. Tailoring retention initiatives to cater to specific job satisfaction levels, tenure-based incentives, and gender-specific policies can significantly mitigate attrition risks. Policy refinements, especially in travel policies and departmental cohesion, coupled with leveraging diversity for an inclusive workplace, stand as pivotal areas for organizational focus.

In conclusion, our data-driven approach furnishes actionable insights imperative for the evolution of effective HR strategies. By leveraging these insights and implementing adaptive policies, organizations can foster an inclusive, supportive work environment, curtail attrition risks, and chart a course toward sustained success.

References

[1] Fernandez, V., & Gallardo-Gallardo, E. (2020). Tackling the HR digitalization challenge: key factors and barriers to HR analytics adoption. Competitiveness Review: An International Business Journal, 31(1), 162-187.

[2] Van den Heuvel, S., & Bondarouk, T. (2017). The rise (and fall?) of HR analytics: A study into the future application, value, structure, and system support. Journal of Organizational Effectiveness: People and Performance, 4(2), 157-178.

[3] M. K. Andersen, "Human capital analytics: the winding road," Journal of Organizational Effectiveness: People and Performance, vol. 4, no. 2, pp. 133–136, 2017.

[4] M. Cao, R. Chychyla, and T. Stewart, "Big Data Analytics in Financial Statement Audits," Accounting Horizons, vol. 29, no. 2, pp. 423–429, 2015.

[5] A. A. Fink and M. C. Sturman, "Hr metrics and talent analytics," The Oxford Handbook of Talent Management, pp. 375–390, 2017.

[6] B. Gaur, V. K. Shukla, and A. Verma, "Strengthening people analytics through wearable IoT device for real-time data collection," in 2019 International Conference on Automation, Computational and Technology Management (ICACTM), IEEE, 2019.