August 8, 2023

Paper on Understanding Healthcare Work Attrition

Abstract:

The purpose of this Python project is to investigate the factors that contribute to healthcare worker attrition in a certain healthcare facility or a subset of institutions. Healthcare worker attrition, often known as employee turnover, is a crucial issue that affects the stability, continuity of care, and general effectiveness of healthcare organizations. In this paper, we hope to look deeper into the topic of healthcare worker attrition and its influence on the healthcare business. Using machine learning techniques and Python programming, we want to create a predictive model that can forecast attrition risks based on historical data and numerous important factors. We believe that our research will provide helpful insights and recommendations to healthcare organizations, allowing them to reduce attrition, improve employee happiness, and eventually improve the overall quality of healthcare service.

SECTION I
Introduction:

Healthcare worker attrition, often known as staff turnover, is a serious issue for healthcare organizations worldwide. It refers to the occurrence of healthcare workers willingly or involuntarily quitting their professions, resulting in a loss of important labour within the healthcare system. Healthcare worker attrition is a complicated issue with far-reaching consequences, impacting not just the stability and efficiency of healthcare organizations, but also the quality of patient treatment and overall healthcare outcomes.

Understanding the root causes of healthcare worker turnover is critical for administrators, politicians, and academics. Identifying the primary variables that lead to staff turnover allows healthcare organizations to create tailored strategies and interventions to address the core causes and increase retention rates. Furthermore, the use of data-driven methodologies, such as machine learning models, may assist forecast attrition concerns, allowing organizations to take proactive efforts to retain and stabilize their personnel.

For example, according to a 2022 study, turnover rates in the healthcare business varied from 19.5% in hospitals to 65% for at-home care providers to 94% in nursing homes. This high turnover place a significant financial and logistical load on healthcare providers. While COVID-19 added to the strain on the healthcare workforce, and the sector will certainly suffer the repercussions for years to come, the healthcare staffing issue existed long before the epidemic.

A. Healthcare worker attrition in the current situation:

In recent years, the healthcare business has seen an increase in staff turnover, owing to a mix of reasons such as growing work expectations, high-stress levels, burnout, restricted career progression prospects, and competitive job markets. The departure of talented and experienced healthcare workers not only affects the continuity of treatment for patients but also places an additional burden on the remaining workforce, potentially reducing job satisfaction and employee morale.

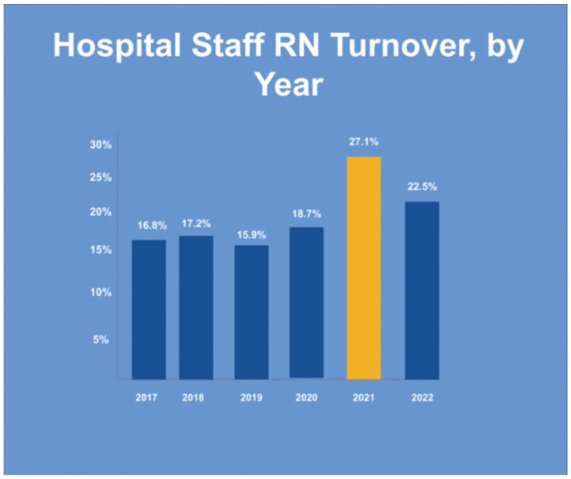

Fig 1: Hospital staff Turnover by year

The healthcare business has been suffering considerable issues in terms of healthcare worker attrition as of my latest report in September 2021. However, the circumstances may have changed since then, and considering the present context is critical for a more accurate perspective. In the current environment, some of the primary variables impacting healthcare worker turnover include:

- Global health crises
- Staff shortage
- Stress and burnout
- Work-life balance
- Compensation and benefits
- Job satisfaction

Understanding the present environment of healthcare worker turnover is critical for healthcare executives and policymakers to establish focused retention and support initiatives. Data-driven techniques, such as machine learning models, can aid in predicting attrition concerns and directing proactive actions. Healthcare organizations can establish a more sustainable and resilient workforce by tackling the fundamental causes of healthcare worker turnover, thereby benefiting both their staff and the patients they serve.

SECTION II
Purpose of the study:

The purpose of this Python project is to investigate the factors that contribute to healthcare worker attrition in a certain healthcare facility or a subset of institutions. Healthcare worker attrition, often known as employee turnover, is a crucial issue that affects the stability, continuity of care, and general effectiveness of healthcare organizations. The project aims to develop a predictive model that can identify potential areas of concern, assisting healthcare administrators and policymakers in implementing targeted strategies for improving retention and job satisfaction among healthcare workers, as well as assisting healthcare institutions in proactively identifying and mitigating attrition risks. This will enable them to implement targeted actions and policies in order to retain their precious personnel, improve patient care, and ensure organizational stability.

In this paper, we hope to look deeper into the topic of healthcare worker attrition and its influence on the healthcare business. Using machine learning techniques and Python programming, we want to create a predictive model that can forecast attrition risks based on historical data and numerous important factors. We believe that our research will provide helpful insights and recommendations to healthcare organizations, allowing them to reduce attrition, improve employee happiness, and eventually improve the overall quality of healthcare service.

SECTION III
Methodology

A. Review strategy:

Key Requirements for the Data Product:

The data product should:
- Predict healthcare worker attrition with high accuracy.
- Provide real-time predictions and insights to enable proactive retention strategies.
- Comply with relevant data privacy and security regulations.

Let us investigate the elements that contribute to healthcare worker attrition and develop a prediction model that can aid healthcare organizations in proactively recognizing and minimizing attrition risks.

B. Key steps/Tasks:

1. Data Collection: Gather relevant data from healthcare institutions database, which includes employee demographics, job roles, work hours, department, tenure, salary, performance evaluations, job satisfaction surveys, and any other factors that may be relevant to employee attrition.
2. Data processing: Clean and preprocess the collected data to handle missing values, remove duplicates, encode categorical variables, and perform any necessary feature engineering.
3. Exploratory data analysis (EDA): Conduct a comprehensive EDA to gain insights into the distribution of employee attrition, explore correlations between different features, and identify potential patterns or trends.
4. Feature selection: Employ suitable feature selection techniques to identify the most significant factors affecting attrition. This step will help in reducing the dimensionality of the data and improve the model's efficiency.
5. Model development: Build a predictive model (e.g., logistic regression, decision tree, random forest, or any other appropriate classifier) to predict healthcare worker attrition based on the selected features. Utilise machine learning libraries in Python to train and evaluate the model's performance.

6. Model evaluation: Assess the model's performance using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score) and validate its results on a separate test dataset to ensure generalizability.
7. Interpretation & Insights: Interpret the model's results to identify the most significant factors contributing to healthcare worker attrition. Provide actionable insights to stakeholders and recommend potential strategies for retention.

Step 1: Data cleaning process:

During data cleaning, we found that there were no null and missing values in the dataset. Then we checked for duplicate rows and columns and dropped the duplicate values. As the dataset was collected from Watson's healthcare report, the data was already clean to the extent that it was ready to use for EDA and model selection.
The main steps we had to take for data cleaning and preparation were to convert all the string variables to float variables in order to get the data ready as an appropriate input for binary classification/prediction. We did this using two methods 'Label Encoding' and 'One hot encoding'.

⇒ Encoding
Label encoding was used for the main predicted variable 'Attrition' while one hot encoding was done for categorical data variables such as 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus' and 'OverTime'.

⇒ Principal component Analysis
After conducting label encoding for all the variables, the number of variables in the dataset increased to 51, and to run models, it was important to include only the important variables in the dataset for analysis.

PCA Results:

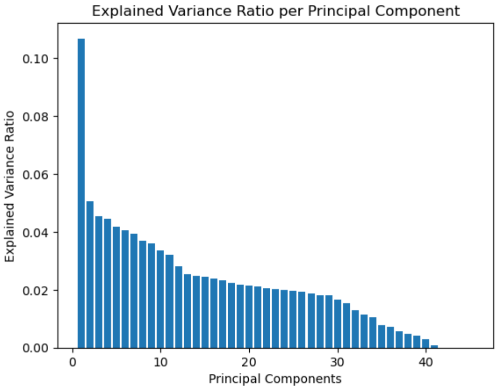

Fig 2: Graph for PCA analysis

We found out that each variable was able to explain a very minute part of the total dataset with only one variable raking in more than 10% of the explained variance ratio and no other variable coming close to explaining 5% of the variability, due to which we ended up using all the variables for our analysis.

Step 2: EDA

Exploratory Data Analysis (EDA) is a critical phase in data analysis that entails graphically and statistically summarising the salient features of a dataset. It entails employing statistical and visual approaches to acquire insights into the structure, patterns, anomalies, and probable correlations between variables in the data. EDA is an important first step before engaging in more complex investigations or modelling.

In short, EDA is a pre-analysis technique that improves the dependability and relevance of later studies. EDA enables researchers and analysts to make educated judgements, draw relevant insights, and build successful models that properly represent the underlying patterns in the data by unravelling the data's story and acquiring a thorough grasp of its properties.

- ○ **Univariate**

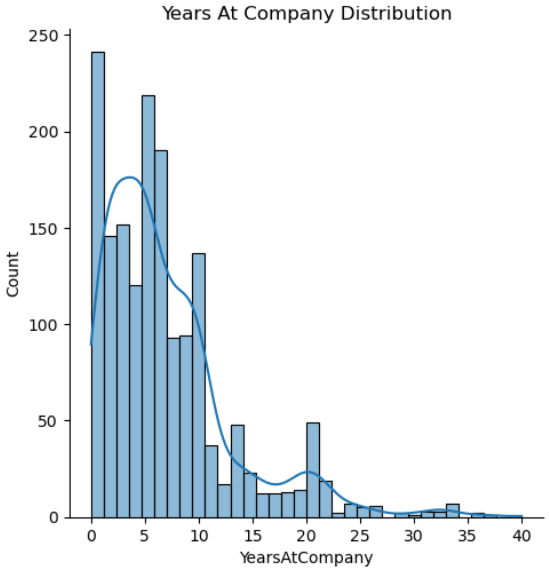

Fig 3: Years at company

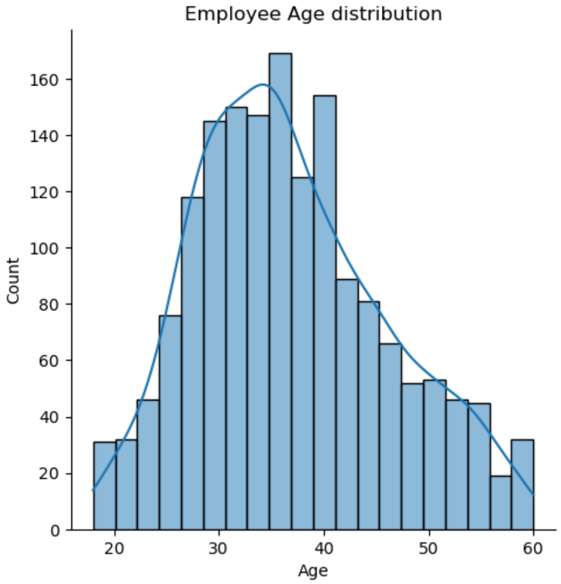

Fig 4: Employee Age Distribution

⇒ Employees on average work at a company for about 7 years

⇒ When looking at the average age of the employees, a bell-curved is formed which makes it an even distribution.

- ○ **Bivariate**

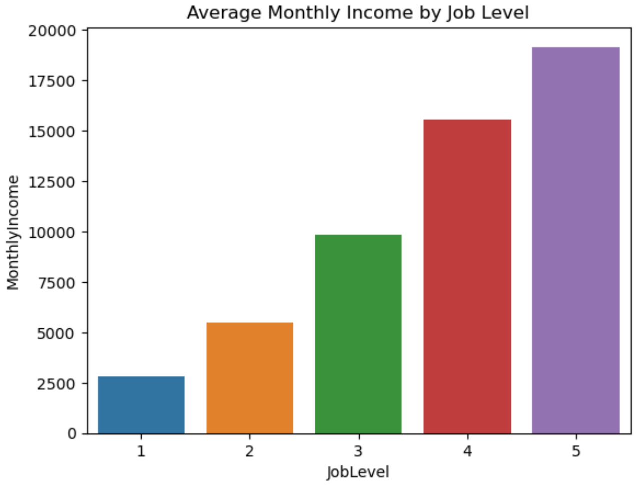

Fig 5: Average monthly income by job level

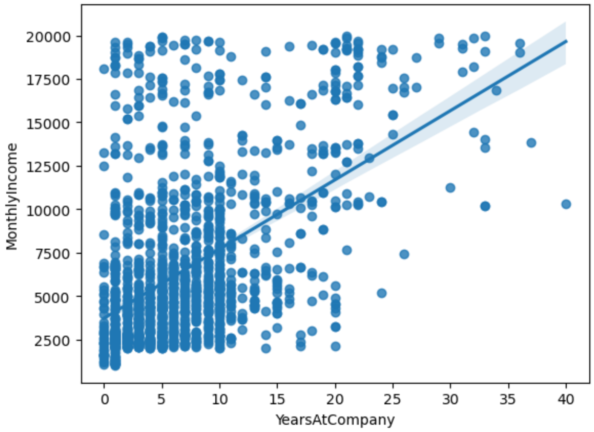

Fig 6: Relation between Monthly income & Years at the company

⇒ Monthly income increases as the job level increases.

⇒ It seems to be a high correlation between the Years at the Company and the Monthly Income. The longer an employee is at the company, the higher is their income.

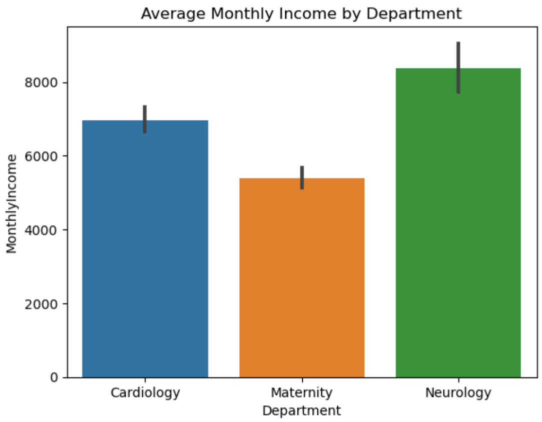

Fig 7: Average monthly income by Department

⇒ The Neurology department seems to have the highest monthly income out of all the departments in the database.

○ **Multivariate**

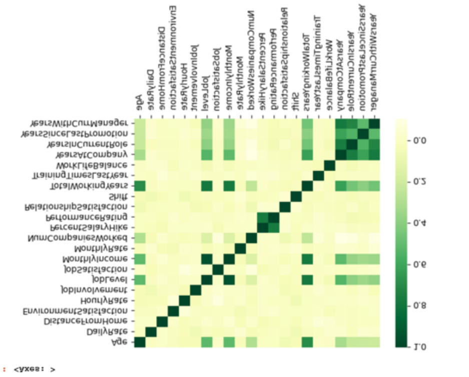

Fig 8: Correlation heat map

⇒ There seems to be a high correlation between Job level and the monthly income of the employees as well as between total years working and monthly income.

⇒ There is also a high correlation between performance rating and yearly salary hike.

Step 4: Model Selection

For supervised learning in healthcare worker attrition prediction, you have a labelled dataset where you already know the attrition status (target variable) for each employee. This allows you to train a model to learn the patterns and relationships between various features (predictor variables) and the attrition outcome.

1. Logistic Regression: For binary classification tasks such as attrition prediction, logistic regression is a popular solution. It's simple, easy to understand, and can predict the likelihood of an employee leaving based on a variety of factors.
2. Decision Trees and Random Forest: Non-linear correlations between attributes and attrition outcomes can be captured using decision trees. Random Forest, a decision tree ensemble, can give more accuracy and generalization.
3. Gradient Boosting Machines (GBM): GBM is a technique for combining numerous weak learners to generate a powerful prediction model. It handles relationships between features effectively and performs well on complicated datasets.
4. Ensemble Methods: To increase prediction performance, ensemble approaches such as AdaBoost and XGBoost may be used to merge numerous models.

Step 5: Model Building

We did a comprehensive analysis of various machine learning models used to predict employee attrition in the healthcare sector. Logistic Regression, k-Nearest Neighbors (KNN), XGBoost, AdaBoost, and Random Forest algorithms were employed and evaluated using multiple metrics including mean squared error (MSE), root mean squared error (RMSE), classification report, area under the curve (AUC), and ROC curves. The objective of this study is to identify the optimum model for accurately predicting attrition among healthcare employees.

C. Results & Insights

The results of the experimental analysis:

TABLE I
Experimental Analysis

	Logistic Regression	KNN	XGBOOST	Random Forest	ADA Boost
RMSE	0.3028	0.135	0.3028	0.1443	0.3345
MSE	0.0874	-0.077	0.0874	0.1073	0.0834
Precision	O.75	0.69	O.75	0.88	0.80

Recall	0.61	0.15	0.61	0.31	0.55
F1-Score	0.67	0.24	0.67	0.46	0.67
AUC	0.78	0.56	0.79	0.65	0.78

D. Conclusion

In this study, we conducted an in-depth comparative analysis of various machine learning models for predicting employee attrition in the healthcare sector. Through a meticulous evaluation of multiple metrics, including mean squared error, root mean squared error, classification report metrics, area under the curve, and ROC curves, we aimed to identify the optimal model.

Among the models tested, AdaBoost emerged as the standout choice for predicting healthcare employee attrition. With a precision of 0.80, AdaBoost showcases a remarkable ability to accurately identify and classify instances of attrition. This high precision is of paramount importance for healthcare institutions seeking to implement proactive measures to mitigate attrition's impact on operations.

Moreover, while AdaBoost excels in precision, XGBoost boasts the highest recall at 0.61. This signifies its capability to effectively identify true positives, indicating its potential in capturing a larger proportion of actual attrition cases. The comparable AUC values for both AdaBoost and XGBoost further underline their comparable overall performance in predictive accuracy.

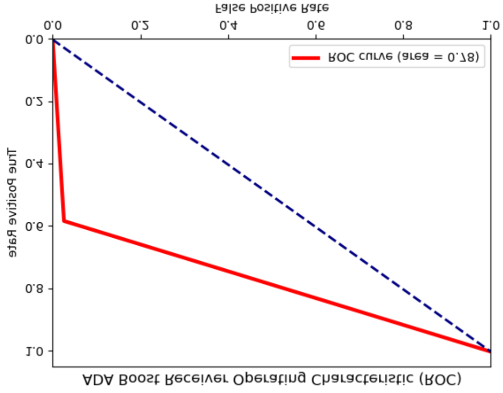

Fig 9: ROC for ADA Boost

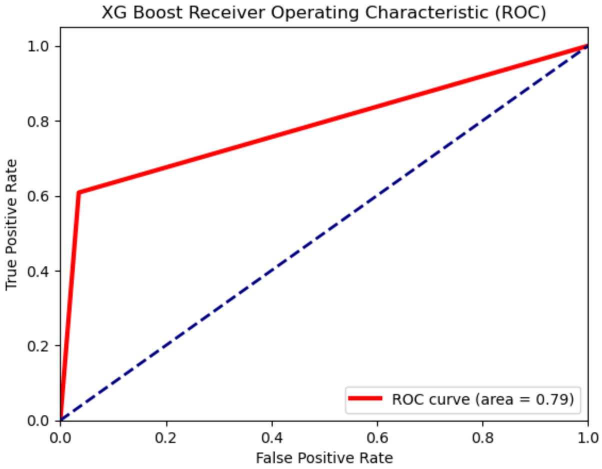

Fig 10: ROC for XG Boost

The F1-scores for both models remain consistent, indicating a harmonious balance between precision and recall. This equilibrium is essential for maintaining a robust and reliable predictive model that can suitably assist healthcare administrators in their decision-making processes related to workforce management.

Furthermore, the slightly lower mean squared error (0.0834) exhibited by AdaBoost in comparison to XGBoost (0.0874) reinforces its superior predictive prowess in this specific context.

In conclusion, considering the collective evidence provided by the evaluation metrics, AdaBoost stands as the optimal model for predicting employee attrition in the healthcare sector. Its impressive precision, combined with the balanced recall, indicates that AdaBoost can effectively aid healthcare institutions in forecasting and addressing workforce attrition. This study highlights the potential of machine learning models in contributing to strategic human resource management within the healthcare domain and paves the way for further research and practical implementation.