

**Artificial Intelligence – IFACET , IITK**

15/07/2024

Ranjeet Kulkarni | Indian Institute of Information Technology Allahabad | IIT2023064

Case Study: HR Attrition

# Problem:

The dataset aims to answer key questions regarding employee attrition:

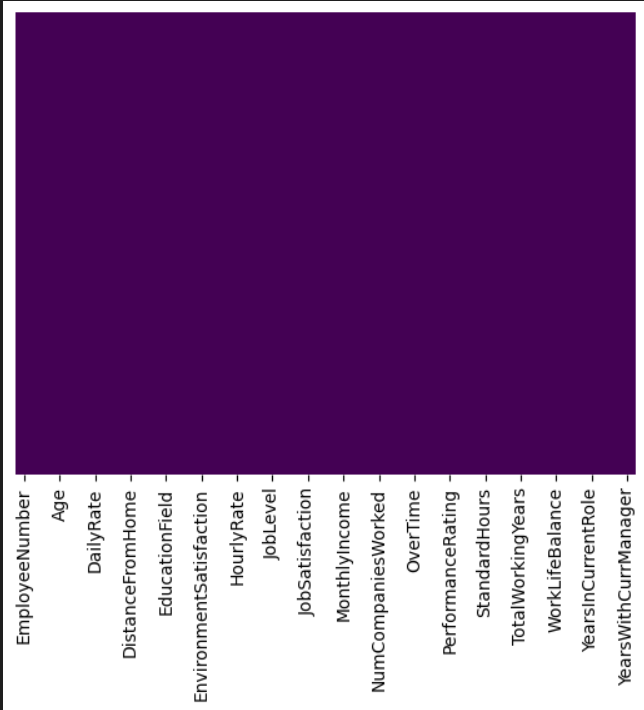
* What factors contribute to identifying attriting employees?
* Can we build a predictive model for employee attrition?
* What metrics are suitable for evaluating such a model?

Provided Data: HR Attrition Data

* EmployeeNumber: Employee Identifier
* Attrition: Did the employee attrite?
* Age: Age of the employee
* BusinessTravel: Travel commitments for the job
* DailyRate: Data description not available
* Department: Employee Department
* DistanceFromHome: Distance from work to home (in km)
* Education: 1-Below College, 2-College, 3-Bachelor, 4-Master, 5-Doctor
* EducationField: Field of Education
* EmployeeCount: Employee Count in a row
* EnvironmentSatisfaction: 1-Low, 2-Medium, 3-High, 4-Very High
* Gender: Employee's gender
* HourlyRate: Data description not available
* JobInvolvement: 1-Low, 2-Medium, 3-High, 4-Very High
* JobLevel: Level of job (1 to 5)
* JobRole: Job Roles
* JobSatisfaction: 1-Low, 2-Medium, 3-High, 4-Very High
* MaritalStatus: Marital Status
* MonthlyIncome: Monthly Salary
* MonthlyRate: Data description not available
* NumCompaniesWorked: Number of companies worked at
* Over18: Over 18 years of age?
* OverTime: Overtime?
* PercentSalaryHike: The percentage increase in salary last year
* PerformanceRating: 1-Low, 2-Good, 3-Excellent, 4-Outstanding
* RelationshipSatisfaction: 1-Low, 2-Medium, 3-High, 4-Very High
* StandardHours: Standard Hours
* StockOptionLevel: Stock Option Level
* TotalWorkingYears: Total years worked
* TrainingTimesLastYear: Number of training attended last year
* WorkLifeBalance: 1-Low, 2-Good, 3-Excellent, 4-Outstanding
* YearsAtCompany: Years at Company
* YearsInCurrentRole: Years in the current role
* YearsSinceLastPromotion: Years since the last promotion
* YearsWithCurrManager: Years with the current manager

General Steps To Be Followed for Any ML Project:

1. **Import Libraries and Load Dataset:**
   * Import necessary Python libraries such as Pandas, NumPy, Matplotlib/Seaborn for data manipulation, analysis, and visualization.
   * Load your dataset into a Pandas DataFrame using pd.read\_csv() or other appropriate methods.
2. **Overview of Data:**
   * Understand the structure of your dataset using methods like head(), info(), describe() to get a glimpse of the data, its columns, data types, and summary statistics.
   * Check for missing values and handle them appropriately if necessary (none here)



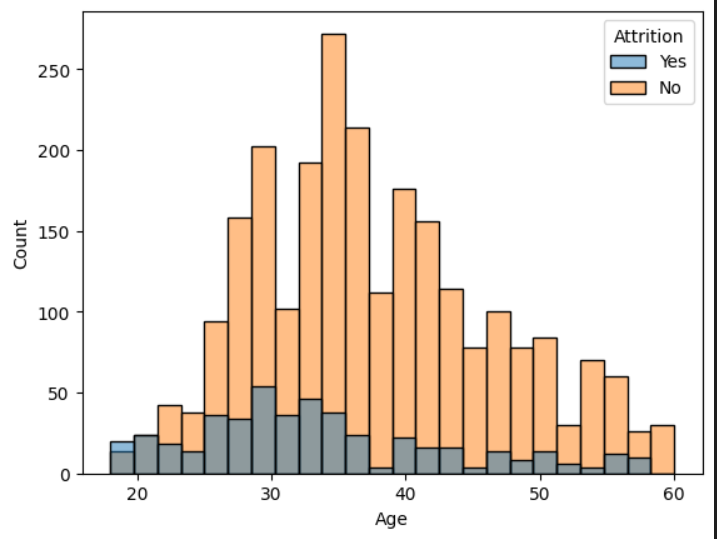
1. **Data Visualization:**
   * Visualize your data to gain insights using plots such as histograms, bar charts, box plots, and scatter plots.
   * Explore relationships between variables to understand correlations and potential patterns.
2. **Data Preparation:**
   * Preprocess your data for machine learning models.
   * Handle categorical variables using techniques like one-hot encoding or label encoding.
   * Scale numerical features if required using StandardScaler or MinMaxScaler.
   * Split your dataset into training and testing sets using train\_test\_split().
3. **Model Building and Evaluation:**
   * Choose a suitable machine learning model based on your problem (e.g., classification for attrition prediction).
   * Train the model on the training data using fit() method.
   * Evaluate the model's performance on the test data using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
   * Tune hyperparameters if necessary using techniques like GridSearchCV or RandomizedSearchCV to optimize model performance.
4. **Conclusion:**
   * Summarize your findings based on model performance and insights gained from data analysis and visualization.
   * Discuss implications for the problem you are addressing (e.g., factors contributing to attrition).
   * Reflect on any limitations of your analysis and propose next steps for further improvement.

Steps followed with respect to model data:

1. Data Analysis
2. Feature engineering
3. Feature selection
4. Model building
5. Model deployment

**Data Analysis:**

* Understanding the type of problem – Supervised or Unsupervised and deciding approach.
* Finding missing values.
* Separating the data into numerical features, categorical features, and year\_feature for analysis.
* Differentiating numerical features into continuous and discrete categories.
* Using Seaborn count plot for each variable in discrete features to analyze trends with respect to attrition.
* Dropping columns that do not provide significant value.
* Using histograms to analyze continuous features and understand their distributions.
* Plotting data points to assess the need for normalization.
* Utilizing box plots to detect outliers in continuous features.
* Analyzing outliers and applying appropriate data manipulation techniques, such as removing data points with low significance.
* Identifying unique subcategories within each categorical feature to determine if feature selection is necessary for specific columns.
* Visualizing categorical features through count plots and deriving insights from the visualizations.
* Understanding the relationships between dependent and independent features to inform model building.



**Model Building:**

**I. Traditional Methods:**

* Splitting the data into train and test sets using an 80% train and 20% test division.
* Utilizing hyperparameter tuning and GridSearchCV to identify the most suitable ML algorithms for building the model.
* Building models using the two most prominent algorithms identified.
* Predicting on the test data for all selected models.
* Obtaining accuracy scores for each selected model to verify the accuracy and error values obtained from GridSearchCV.

**II. Neural Networks Approach:**

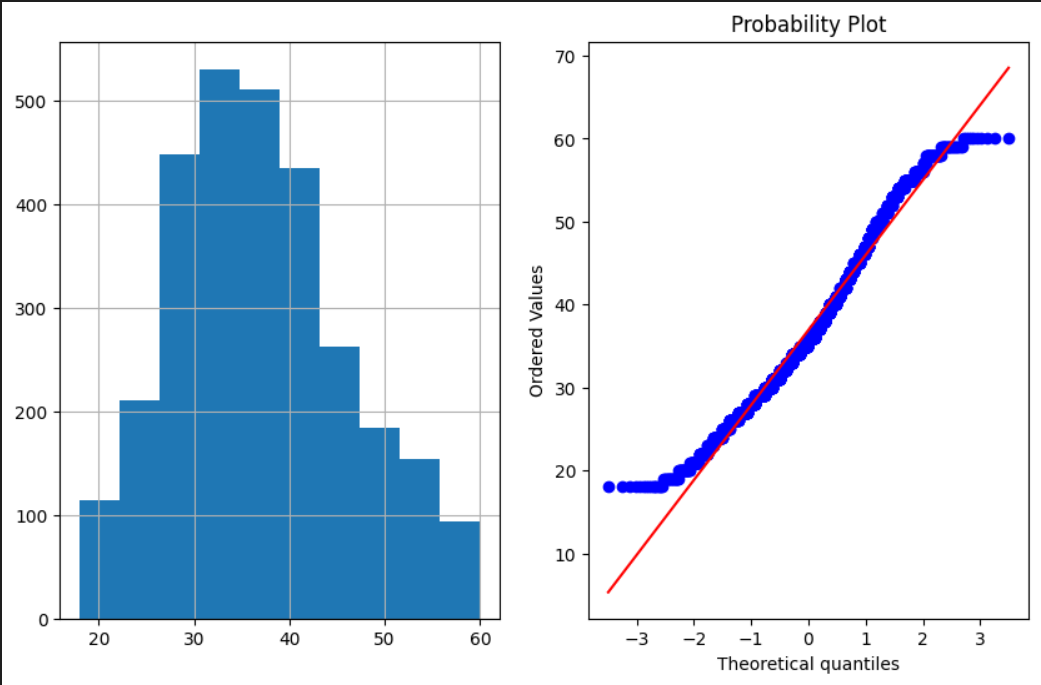
* Installing the TensorFlow library for neural network implementation.
* Normalizing the input data using MinMaxScaler suitable for neural networks.
* Dividing the data into train and test sets.
* Fitting the neural network model on the train dataset with a general epoch value of 100.
* Calculating the score for both the train and test data to evaluate model performance.
* Assessing the overall accuracy achieved by this approach.

**Model Deployment:**

1. **Creating Pickle File:**
   * Create a pickle file to save the trained model and the scaling object (e.g., StandardScaler) using Python's pickle module.
2. **Setting Up New Environment:**
   * Set up a new Python environment using VS Code or any preferred IDE.
3. **Installing Necessary Libraries:**
   * Install all required libraries for your Flask application, including Flask itself and any dependencies your model needs to run.
4. **Pushing Files to GitHub Repository:**
   * Push your entire project, including your trained model pickle file, app.py, and any necessary files (like HTML templates and CSS) to a GitHub repository for version control and deployment.
5. **Creating app.py:**
   * Implement the Flask application (app.py) that serves as the endpoint for model prediction:
     + Obtain data from an HTML webpage in JSON format using Flask's request module.
     + Convert the received JSON data into a pandas DataFrame.
     + Apply encoding and scaling transformations based on the preprocessing steps identified during data analysis.
     + Use the trained model (loaded from the pickle file) to predict outputs for the input values.
     + Display the predicted output on the webpage or return it as a JSON response to the user.
6. **Deployment Considerations:**
   * Note that the model was not deployed on any cloud platform due to scale and time limitations, but this setup can be adapted for deployment on platforms like Heroku, AWS, or Azure as needed.

Observations:

1. Higher levels of education are associated with lower attrition rates.
2. Employees with higher levels of Environment Satisfaction tend to experience lower attrition.
3. Greater job involvement correlates with decreased attrition rates.
4. Higher job levels are linked to lower attrition rates.
5. Increased job satisfaction is associated with lower attrition rates.
6. Employees who have worked in five or more companies show higher attrition rates.
7. Higher percentage salary hikes correlate with higher attrition rates.
8. Employee count does not affect attrition and was disregarded.
9. Standard hours have no direct impact on attrition.
10. Distance from home and monthly income exhibit skewed distributions.
11. All other continuous features show Platykurtic distributions (flat peak).
12. Probability plots for continuous features indicate values closely aligned with the straight line, suggesting the data is not extreme.



1. The dataset does not contain significant outliers in continuous features.
2. Outliers were detected in Daily Rate and Employee Number columns; however, Employee Number is irrelevant and was subsequently dropped from analysis.

The feature is BusinessTravel and number of categories are 3

The feature is Department and number of categories are 3

The feature is EducationField and number of categories are 6

The feature is Gender and number of categories are 2

The feature is JobRole and number of categories are 9

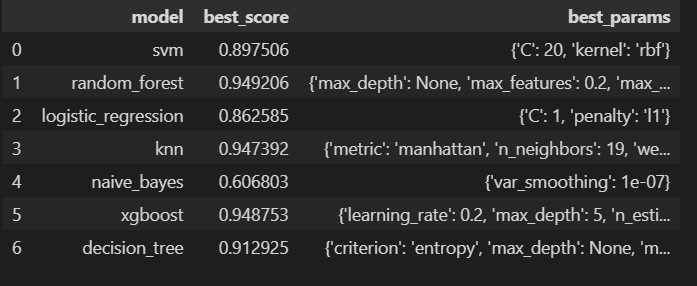
The feature is MaritalStatus and number of categories are 3

The feature is Over18 and number of categories are 1The feature is OverTime and number of categories are 2.

1. Hence none of categorical features have too many unique categories.

Thus its easy to make sure model is not overfitted

1. No attrition corresponds to about 2400 instances, while attrition cases are less than 500, indicating the need for oversampling due to the relatively small data size.
2. Employees who travel frequently show a higher percentage of attrition.
3. The majority of employees are from the Research and Development department, which experiences the highest attrition rates.
4. There are approximately 1000 female employees with less than 200 experiencing attrition, compared to about 1400 male employees with around 250 experiencing attrition, suggesting a higher attrition rate among females.
5. Singles are most likely to change companies, which aligns with the observation regarding age. Employees aged between 25 to 32-33 exhibit the highest attrition rates.
6. Overtime work correlates significantly with attrition, indicating it as a major contributing factor.
7. During hyperparameter tuning, the model was refined to improve performance.



1. **KNeighborsClassifier:** Using Manhattan distance metric with 19 neighbors and weighted by distance.
2. **Accuracy on KNN:** Achieved 95.1%.
3. **RandomForestClassifier:** Utilizing 120 estimators with 20% maximum features and full sampling.
4. **Accuracy on Random Forest:** Achieved 95.64%.
5. **Neural Networks Approach:**
6. **On Test Data:** Achieved an accuracy of 89.97%.
7. **On Train Data:** Achieved an accuracy of 95.41%.

### 

### Conclusions of the Attrition Prediction Analysis

**Summary of Observations and Findings:**

1. **Education and Attrition:**
   * Higher levels of education are associated with lower attrition rates. Employees with advanced degrees tend to stay longer in their positions, possibly due to better job opportunities and satisfaction derived from their roles.
2. **Environment Satisfaction:**
   * Employees with higher levels of environment satisfaction experience lower attrition. This suggests that improving workplace conditions and employee satisfaction can significantly reduce turnover.
3. **Job Involvement:**
   * Greater job involvement correlates with decreased attrition rates. Employees who are more engaged and involved in their work are less likely to leave their jobs.
4. **Job Level:**
   * Higher job levels are linked to lower attrition rates. Senior employees tend to have more stability and satisfaction, leading to longer tenure.
5. **Job Satisfaction:**
   * Increased job satisfaction is associated with lower attrition rates. Ensuring that employees find their work fulfilling and rewarding can help retain talent.
6. **Experience in Multiple Companies:**
   * Employees who have worked in five or more companies show higher attrition rates. This may indicate a trend of job-hopping among these individuals.
7. **Salary Hikes:**
   * Higher percentage salary hikes correlate with higher attrition rates. This counterintuitive finding suggests that substantial salary increases might be given to retain employees who are already considering leaving, or that such increases could raise expectations and lead to dissatisfaction if not met.
8. **Employee Count and Standard Hours:**
   * Employee count does not affect attrition and was disregarded. Similarly, standard hours have no direct impact on attrition, indicating that other factors are more significant.
9. **Distribution of Continuous Features:**
   * Distance from home and monthly income exhibit skewed distributions, while all other continuous features show Platykurtic distributions (flat peak). This suggests that most employees have similar values in these features, with few extreme values.

 **Normalization and Outliers:**

* Probability plots for continuous features indicate values closely aligned with the straight line, suggesting the data is not extreme. The dataset does not contain significant outliers in continuous features. Outliers were detected in Daily Rate and Employee Number columns; however, Employee Number was irrelevant and dropped from analysis.

 **Categorical Features:**

* The dataset includes several categorical features with a manageable number of unique categories:
  + Attrition: 2 categories
  + BusinessTravel: 3 categories
  + Department: 3 categories
  + EducationField: 6 categories
  + Gender: 2 categories
  + JobRole: 9 categories
  + MaritalStatus: 3 categories
  + Over18: 1 category
  + OverTime: 2 categories

1. **Imbalance in Attrition Data:**
   * There is a significant imbalance in the dataset, with about 2400 instances of no attrition and less than 500 instances of attrition. This imbalance necessitates the use of oversampling techniques like SMOTE to ensure balanced training data.
2. **BusinessTravel and Attrition:**
   * Employees who travel frequently show a higher percentage of attrition, indicating that frequent travel might lead to job dissatisfaction and turnover.
3. **Departmental Trends:**
   * The majority of employees are from the Research and Development department, which also experiences the highest attrition rates. Targeted interventions in this department could help reduce turnover.
4. **Gender and Attrition:**
   * There are approximately 1000 female employees with less than 200 experiencing attrition, compared to about 1400 male employees with around 250 experiencing attrition. This suggests a higher attrition rate among females, indicating potential gender-specific issues that need addressing.
5. **Marital Status and Age:**
   * Single employees are most likely to change companies, which aligns with the observation that employees aged between 25 to 32-33 exhibit the highest attrition rates. This demographic may be more inclined to seek new opportunities.
6. **Overtime:**
   * Overtime work correlates significantly with attrition, indicating it as a major contributing factor. Reducing overtime or compensating it adequately could help in retaining employees.

**Model Performance:**

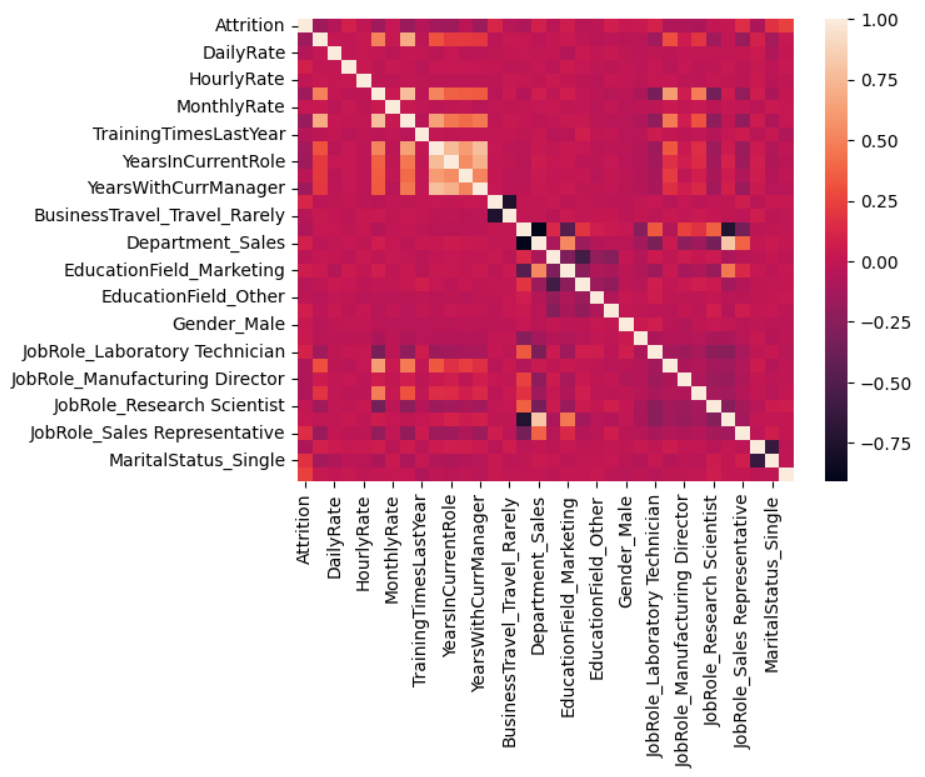
1. **KNeighborsClassifier:**
   * Configuration: Manhattan distance metric with 19 neighbors, weighted by distance.
   * Accuracy: 95.1%.
2. **RandomForestClassifier:**
   * Configuration: 120 estimators with 20% maximum features and full sampling.
   * Accuracy: 95.64%.
3. **Neural Networks Approach:**
   * Configuration: Normalized input data, general epoch value of 100.
   * Accuracy on Test Data: 89.97%.
   * Accuracy on Train Data: 95.41%.

Concluding Remarks:

The analysis reveals several key factors influencing employee attrition, including education level, job satisfaction, job involvement, distance from home, and job role. The models developed, particularly the Random Forest and KNN classifiers, show high accuracy, demonstrating their effectiveness in predicting attrition. However, the neural network model, while effective, showed slightly lower accuracy on test data, indicating potential for further refinement.

Addressing the identified factors, such as reducing overtime and improving job satisfaction, could help mitigate attrition. Additionally, targeted interventions for demographics with higher attrition rates, like young and single employees, as well as those in the Research and Development department, could further enhance employee retention strategies.

Overall, the comprehensive approach combining data analysis, feature engineering, and robust model development has provided valuable insights and actionable recommendations for managing employee attrition.



Important Graphs

