Attrition Analysis and Prediction
Mid-Semester Project Report | Group 11 (ML Sec - A)

Varun Bharti
IIIT Delhi
varun22562@iiitd.ac.in

Yash Sinha
IIIT Delhi
yash22590@iiitd.ac.in

Shashwat Jha
IIIT Delhi
shashwat22472@iiitd.ac.in

Vatsal Gupta
IIIT Delhi
vatsal22564@iiitd.ac.in

Vaibhav Sehra
IIIT Delhi
vaibhav22554@iiitd.ac.in

Abstract

Employee attrition is a critical concern for organizations, as it affects both operational efficiency and employee morale. This study analyzes employee attrition by applying machine learning models to the IBM HR Analytics Employee Attrition dataset. After conducting comprehensive data preprocessing, including feature scaling and label encoding, we implemented five classification models: Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, and Support Vector Classifier (SVC). Exploratory data analysis revealed key insights into correlations between employee characteristics and attrition rates. Our models were evaluated based on accuracy, precision, recall, and F1-score. Logistic Regression achieved the highest accuracy at 89.11%, while Random Forest and SVC also showed competitive results. These findings highlight the importance of feature selection and model choice in predicting employee turnover, with implications for human resource management and retention strategies

1. Motivation

Employee attrition is a significant concern for organizations, affecting productivity, morale, and overall performance. Studies show that understanding the factors contributing to attrition can help companies retain talent and reduce turnover costs. This analysis seeks to empower organizations to improve employee retention and performance.

2. Introduction

Employee attrition is a pressing issue faced by many organizations, with significant consequences on productivity, operational costs, and overall company performance. Understanding the factors contributing to employee attrition can empower companies to develop effective retention strategies. The rise of data analytics has provided organizations with an opportunity to predict attrition by leveraging machine learning algorithms on historical employee data.

In this study, we utilized the IBM HR Analytics Employee Attrition dataset to predict which employees are at risk of leaving the company. This dataset contains information on various attributes of employees, such as job role, salary, tenure, and work environment, making it suitable for classification tasks.

The methodology adopted includes a detailed data preprocessing phase involving the removal of redundant features, handling class imbalances, and normalizing features using Min-Max scaling. We employed exploratory data analysis to identify significant correlations between employee attributes and attrition. This was visualized through various plots allowing us to gain valuable insights into the factors contributing to attrition.

To predict attrition, we applied five classification algorithms: Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, and Support Vector Classifier (SVC). These models were evaluated based on key performance metrics, including accuracy, precision, recall, and F1-score. The results highlight that Logistic Regression outperformed other models with an accuracy of 89.11%, followed closely by Random Forest and SVC.

3. Literary Survey

We looked at 3 different papers, each looking at the same topic using different methods. From these, we hoped to identify which method would give us the best accuracy along with an expected accuracy.

3.1. Employee Attrition: Analysis of Data Driven Models [2]

The paper compares the performance of various Machine Learning, Ensemble Machine Learning and Deep Learning Algorithms on the IBM Human Resource Analytics Performance dataset. The authors first performed feature scaling using Min-Max scaling and removed any null values from the dataset.Then they did feature selection using Principal Component Analysis (PCA) to reduce the dimensionality of the dataset. Afterwords, they employed Label Encoding to obtain a numerical representation of the categorical features. Following this, they split the data into 75% for training and 25% for testing to evaluate model performance.

After the data pre-processing was done, they trained various models on the data and compared their effectiveness on the basis of performance metrics like accuracy, precision, recall and F1 score. It was found that linear models like Logistic Regression outperformed the other models. Apart from this, the Deep Learning Models, particularly FNNs performed exceptionally well on the data.

3.2. Employee Attrition Prediction [4]

In this study, conducted by a team from PESIT-BSC, Bangalore, the authors try different models to ascertain the likelihood of an employee leaving a company. The model that gives the best accuracy primarily employs the k-Nearest Neighbors (KNN) algorithm, leveraging key features such as employee performance evaluations, monthly work hours, and tenure with the company. The research differentiates itself by also exploring alternative machine learning methods, including ANNs like MLP, Decision Trees, and Logistic Regression for comparison.

The dataset is categorized under binary class labels indicating whether employees have left the company. After converting all categorical data into numerical values and scaling, the dataset underwent a 70:30 split for training and testing, respectively. For KNN, the distance from neighbors is calculated using Manhattan distance, and then the class is decided by a majority vote. Model performance was evaluated based on metrics such as AUC, accuracy, and F1 score, with the KNN model demonstrating superior performance with an accuracy of 94.32Intuitively, data points that are close to each other are likely to have the same outcome of attrition, which is the basis for choosing the KNN algorithm. The results of this research showed the superiority of the KNN classifier in terms of accuracy and predictive effectiveness, by means of the ROC curve.

3.3. Employee's Attrition Prediction Using Machine Learning Approaches [3]

This research looks into the effectiveness of different machine learning approaches for employee attrition prediction. Their study starts by tackling the data imbalance challenge, by using enhanced training processes and proper feature preprocessing. The authors used five different machine learning algorithms: Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, and Logistic Regression. They evaluated model performance using multiple relevant metrics including precision, recall, F1-score, and AUC-ROC curves. They moved beyond traditional accuracy measures which were shown to be not as effective for datasets that are imbalanced. The results showed that Logistic Regression achieved superior performance with 86% accuracy and precision and recall of (0.87) & (0.86) respectively. The study also revealed that certain features like job title had more significant impact on attrition prediction than conventional factors like age. These findings suggest that Traditional accuracy metrics are not as reliable and simpler linear models can be effective for attrition prediction when properly implemented.

4. Dataset

The dataset [1] consists of **1,470 entries** and **35 features**. After preprocessing and performing EDA, it contains **1,470 entries** and **31 features**, with no missing (NaN) values detected.

To gain insights from the dataset, a comprehensive exploratory data analysis (EDA) was conducted. This involved visualizing the data through various plots, including **11 different plots** and **over 30 graphs** which included histograms, box plots, heatmaps, violin plots, pie plot, line graph, bar graph, area chart, spline, dot plot, bidirectional bar chart.These visualizations guided the interpretation of patterns in the data and informed subsequent modeling decisions. From these observations , following can be inferred:

4.1. Correlations

The strongest positive correlations with the target variable (employee attrition) include:

- Performance Rating
- Monthly Rate
- Num Companies Worked
- Distance From Home

The strongest negative correlations with the target variable include:

- Total Working Years
- Job Level
- Years In Current Role
- Monthly Income

4.2. Imbalance

The dataset is imbalanced, with a majority of observations representing currently active employees. This suggests a need for careful handling during model training to avoid bias towards the majority class.

4.3. Redundant Features

Several features were identified as redundant for the analysis:

- EmployeeCount
- EmployeeNumber
- StandardHours
- Over18

`EmployeeNumber` wasnt significantly correlated while the other three had only 1 unique value i.e demonstrated a linear pattern which wouldnt affect the results in any way or form

4.4. Observations

- **Marital Status**: Single employees exhibit the highest proportion of leavers compared to married and divorced counterparts.
- **Tenure**: Approximately **10%** of leavers departed when reaching their **2-year anniversary** with the company.
- **Loyalty**: Employees with higher salaries and more responsibilities demonstrate a lower proportion of leavers compared to their counterparts.
- **Commute**: Individuals living further from the workplace tend to have a higher proportion of leavers.
- **Travel**: Employees who travel frequently show an increased tendency to leave the organization.
- **Overtime**: Employees required to work overtime have a higher likelihood of leaving compared to those who do not.
- **Job Role**: A significant percentage of leavers are found among Sales Representatives.
- **Previous Employment**: Employees with a history of bouncing between workplaces (multiple previous companies) show a higher likelihood of leaving.
- **Job Level & Income**: Workers with lower job levels, monthly income, years at the company, and total working years are more likely to quit.
- **Business Travel**: Employees who travel frequently are more prone to quitting.
- **Department**: Employees in Research & Development tend to have a lower attrition rate compared to those in other departments.
- **Education Field**: Workers with degrees in Human Resources and Technical fields exhibit higher attrition rates compared to employees from other educational backgrounds.
- **Gender**: Male employees show a higher likelihood of quitting.
- **Job Role**: Laboratory Technicians, Sales Representatives, and Human Resources personnel are more likely to leave compared to other job roles.

5. Methodology

The dataset used in this study was the HR Employee Attrition dataset, which was pre-processed and analyzed through a series of classification models. The steps involved in data preparation, feature engineering, and model training are detailed below.

5.1. Data Preprocessing

The dataset was first loaded using `pandas`, and based on the EDA conducted earlier, four columns were dropped. Categorical columns with fewer than 50 unique values were identified for label encoding. The `Attrition` column, which serves as the target variable, was converted to a binary form using the `category` datatype and encoded into numerical labels using `LabelEncoder`. The remaining categorical columns were similarly label encoded.

5.2. Feature Scaling

Before training, the features were scaled using Min-Max scaling. The feature matrix `X` was normalized such that all feature values fall within the range [0, 1], ensuring that no feature would disproportionately influence the models due to differences in scale.

5.3. Train-Test Split

The dataset was split into training and testing sets using an 80:20 split, with a fixed random seed of 42 to ensure reproducibility. This split helped evaluate model performance on unseen data during testing.

5.4. Model Selection and Training

We implemented multiple classification algorithms to predict employee attrition:

- **Logistic Regression**: A baseline linear model was trained using the `LogisticRegression` class.

- **Decision Tree Classifier**: A decision tree was trained using `DecisionTreeClassifier` to capture non-linear relationships between features.
- **Random Forest Classifier**: A Random Forest model was trained using `RandomForestClassifier`. This ensemble method combines multiple decision trees to improve classification accuracy and control overfitting
- **Gaussian Naive Bayes**: A probabilistic model was trained using `GaussianNB`, leveraging the assumption of feature independence.
- **Support Vector Classifier (SVC)**: Both radial basis function (RBF) and linear kernels were evaluated using `SVC`.

Each model was trained on the normalized training data, and predictions were made on the test set. Classification reports, including precision, recall, F1-score, and accuracy, were generated to compare the performance of each classifier.

6. Results

The primary evaluation metrics used to assess model performance were accuracy and the detailed classification report. These metrics provided insight into each model's ability to correctly predict employee attrition.

The classification models were evaluated using accuracy as well as detailed statistics from the classification report, including precision, recall, and F1-score. The results for each method are summarized in Table 1.

Method	Accuracy	Precision	Recall
Logistic Regression	89.11	0.88	0.89
Decision Tree	77.21	0.78	0.77
Random Forest	87.75	0.87	0.88
Gaussian Naive Bayes	84.69	0.87	0.85
SVC (RBF kernel)	88.43	X0.90	0.88
SVC (Linear kernel)	88.77	0.88	0.89

Table 1. Summary of classification model results with accuracy and classification report statistics.

This methodology allowed for a comprehensive comparison of different classifiers in predicting employee attrition.

7. Conclusion

7.1. Learnings

In this project, we successfully implemented and compared multiple machine learning models to predict employee attrition. Through exploratory data analysis (EDA) and feature selection, we identified key factors influencing attrition and trained various classifiers such as Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, and Support Vector Machines (SVC). Our learning highlighted the importance of data preprocessing, especially in handling imbalanced datasets, and how different models respond to these preprocessing steps.

7.2. Future Wrok

Despite achieving competitive accuracy with linear models like Logistic Regression and ensemble methods like Random Forest, there is still room for improvement. In our future work, we plan to explore more advanced preprocessing techniques such as Principal Component Analysis (PCA) to further reduce dimensionality and improve model performance. We also aim to experiment with a broader range of models, including deep learning approaches such as Convolutional Neural Networks (CNNs), Fully Connected Neural Networks (FNNs), and Multi-Layer Perceptrons (MLPs). Additionally, building an ensemble pipeline combining multiple models is another avenue for future exploration.

While we achieved significant insights from our EDA and model comparison, refining the dataset based on these insights is a critical step for our next iteration. By tweaking the dataset and experimenting with various feature engineering techniques, we aim to improve our overall model performance.

7.3. Contributions

- **Varun Bharti**: Led the overall coordination of the project. Managed data preprocessing and was responsible for of all models. Handled writing whole report.
- **Yash Sinha**: Was responsible for performing the EDA and getting insights on the dataset. Implemented EDA techniques including plotting all the graphs , preprocessing data and getting the key insights from the same
- **Shashwat Jha**: Was responsible for researching and reading the papers and analysing the existing work done in the domain.
- **Vatsal Gupta**:Was responsible for researching and reading the papers and analysing the existing work done in the domain. Also contributed in performing EDA
- **Vaibhav Sehra**: Contributed to the EDA , as well as responsible for researching and reading the papers and analysing the existing work done in the domain.

Through this collaborative effort, we gained practical experience with various machine learning models and techniques, which will guide our future research and exploration in the domain of employee attrition prediction.

References

[1] Internet. Ibm hr analytics employee attrition performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analyticsattrition-dataset. 2

[2] Divya Sahu Mahima Dogra Manju Nandal, Veena Grover. Employee attrition: Analysis of data driven models, 2020. https://publications.eai.eu/index.php/IoT/article/view/4762/2793. 2

[3] K. K. Mohbey. Employee's attrition prediction using machine learning approaches real-time applications, 2020. DOI: 10.4018/978-1-7998-3095-5.ch005. 2

[4] Rakshit Vahi Rahul Jana Abhilash GV Deepti Kulkarni Rahul Yedida, Rahul Reddy. Employee attrition prediction, 2018. arXiv:1806.10480. 2