Attrition Analysis and Prediction
End-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, we implemented various ML models like Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, Support Vector Classifier, Gradient Boosting etc. Exploratory data analysis revealed key insights into employee characteristics and attrition rates. Our models were evaluated based on accuracy, precision, recall, and F1-score. Random Forest Classifier achieved the highest accuracy at 94.51%, while Gradient Boosting and Logistic regression 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. All the work can be accessed here : GitHub Repository

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. We employed exploratory data analysis to identify significant correlations between employee attributes and attrition. This was visualized through various plots giving valuable insights into the factors contributing to attrition.

To predict attrition, we applied different machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes,Support Vector Classifier (SVC), Gradient Boosting etc. These models were evaluated based on key performance metrics, including accuracy, precision, recall, and F1-score. We implemented these models using different data preprocessing techniques to get possible output

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:

- 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
- 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.
- 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
- Various other observations were also made which has been discussed in prior reports

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.

5.1. Basic Preprocessing and Model Implementation

We employed a basic data preprocessing technique. From the dataset, four columns were dropped based on the EDA conducted earlier. Categorical columns with fewer than 50 unique values were identified for label encoding. Attrition column was converted to a binary form using the `category` datatype and encoded into numerical labels using `LabelEncoder`. Before training, the features were scaled using Min-Max scaling. The dataset was split into training and testing sets using an 80:20 split, with a fixed random seed of 42.

Following this, we implemented multiple classification algorithms to predict employee attrition. These included : Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, and Logistic Regression.

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. The best accuracy achieved was 89.11%.

5.2. Advanced Data Preprocessing

Following the basic model implementation, we advanced to implementation of preprocessing pipeline for the dataset. This processing pipeline consisted of following steps :

- The columns EmployeeCount, EmployeeNumber, Over18, and StandardHours were removed from the dataset
- Features with only a single unique value across all rows were identified and dropped, as they do not contribute to the predictive model.
- The numeric features were scaled to a range of [0, 1] using Min-Max Normalization. This step ensured that all numerical features were on a consistent scale.
- The target variable, Attrition, was converted into a binary numerical format (0 for ”No” and 1 for ”Yes”) using label encoding.
- The dataset exhibited class imbalance between the attrition categories. To address this, Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples of the minority class, ensuring an equal distribution of target classes in the resampled dataset.

5.3. Advanced Model Implementation

Based on the described preprocessing, we implemented five different models. Classification reports, including precision, recall, F1-score, and accuracy, were generated to compare the performance of each model. The best accuracy achieved was 94.51%. These are as follows :

5.3.1 Model 1: XGBoost Classifier with Grid Search Cross-Validation and Feature Importance Analysis

The first model implemented was an XGBoost Classifier, optimized using Grid Search Cross-Validation to identify the best combination of hyperparameters. The parameters tuned included learning rate (0.01, 0.1, 0.2), max depth (3, 5, 7), and n estimators (50, 100, 200), with a 5-fold cross-validation approach ensuring robust evaluation of each parameter combination.

5.3.2 Model 2 : XGBoost Classifier with Bayesian Optimization and Recursive Feature Elimination

The second model implemented was also an XGBoost Classifier, optimized using Bayesian Optimization for hyperparameter tuning and combined with Recursive Feature Elimination (RFE) for feature selection. Bayesian Optimization efficiently tuned key parameters to improve model performance, while RFE reduced the feature set to the most predictive variables, enhancing interpretability and computational efficiency.

5.3.3 Model 3: Hybrid XGBoost and Random Forest with Recursive Feature Elimination and Bayesian Optimization

The third model combines the strengths of XGBoost and Random Forest using a soft-voting ensemble approach for improved predictive performance. Feature selection was performed using Recursive Feature Elimination with Cross-Validation (RFECV), which identified the most important features through a systematic reduction process. The reduced feature set was

then used for training the models, ensuring computational efficiency and interpretability.

5.3.4 Model 4: Stacking Classifier with XGBoost, LightGBM, Random Forest, and Gradient Boosting

The fourth model is a Stacking Classifier, which combines multiple base models to improve prediction accuracy. In this implementation, the base models include XGBoost, LightGBM, Random Forest, and Gradient Boosting. The final predictions are made by a Logistic Regression model that aggregates the predictions of the base models. Prior to training the stacking classifier, Principal Component Analysis was applied to the data to reduce dimensionality while preserving 95% of the variance, which helps in improving computational efficiency and reducing overfitting.

5.3.5 Model 5: TPOT AutoML Classifier and Stacking Classifier with MLP

The fifth model integrates TPOT, with a Stacking Classifier. TPOT uses genetic algorithms to automatically optimize machine learning pipelines, selecting the best model and hyperparameters for the task. After training on the data, the best pipeline was exported and evaluated. Additionally, a Stacking Classifier was implemented, combining XGBoost, LightGBM, and a Multilayer Perceptron (MLP) classifier, with a Logistic Regression model as the final estimator.

6. Results

6.1. Base Models Results

The primary evaluation metrics used to assess model performance were accuracy and the detailed classification report.The results for each method are summarized in Table 1.

Method	Accuracy	Precision	Recall	F1 Score
Logistic R	89.11	0.88	0.89	0.87
Decision Tree	77.21	0.78	0.77	0.78
Random Forest	87.75	0.87	0.88	0.83
Naive Bayes	84.69	0.87	0.85	0.86
SVC (RBF)	88.43	0.90	0.88	0.84
SVC (Linear)	88.77	0.88	0.89	0.85

Table 1. Results for base model implementation

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

6.2. Advanced Models Results

The advanced models implemented has been assessed on accuracy and the detailed classification report.The results for each method are summarized in Table 2.

Method	Accuracy	Precision	Recall	F1 Score
Model 1	92.56	0.93	0.93	0.93
Model 2	93.24	0.93	0.93	0.93
Model 3	94.51	0.94	0.94	0.93
Model 4	92.56	0.93	0.93	0.93
Model 5	92.97	0.93	0.93	0.93

Table 2. Results for advanced model implementation

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. Further, we had a deeper understanding on ensamble models and leveraged the concept of pipelines.

7.2. Contributions

- **Varun Bharti**: Led the overall coordination of the project. Managed data preprocessing and was responsible for of implementation 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. Was responsible for of implementation of all models.
- **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.

– **Vaibhav Sehra**: Was 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] 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.

[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.

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