Report on Employee Attrition Prediction

Name: Bhavik Jain

Mail: bhaviklunkad208@gmail.com

Phone No.: 7225080632

Contents

Introduction

Worker whittling down, the marvel of representatives taking off an organization, could be a critical concern for businesses because it can lead to efficiency misfortunes, expanded enlistment costs, and disturbances in workflow. To address this challenge, businesses frequently depend on prescient analytics to distinguish components contributing to steady loss and create methodologies for maintenance.

In this report, we analyze the IBM HR Analytics Worker Whittling down & Execution dataset to get it the variables impacting worker whittling down and create prescient models to distinguish workers at hazard of taking off the organization. We take after a organized approach, counting dataset investigation, preprocessing, demonstrate improvement, assessment, and optimization.

Dataset Analysis and Preprocessing

The dataset contains different qualities related to representative socioeconomics, work parts, fulfilment levels, execution appraisals, etc., at the side a target variable showing whether an representative has cleared out the company. We perform the taking after preprocessing steps:

- **Handling Lost Values:** We check for lost values within the dataset and handle them suitably, guaranteeing that the information is total for examination.
- **Encoding Categorical Factors:** We encode categorical factors utilizing one-hot encoding to change over them into a arrange reasonable for machine learning calculations.
- **Scaling Numerical Highlights:** We scale numerical highlights to guarantee that they have a comparable run, which can move forward the execution of a few machine learning calculations.

Model Development

Within the demonstrate advancement stage, we center on part the dataset into preparing and testing sets and selecting appropriate machine learning calculations for double classification. The objective is to construct prescient models that can precisely classify whether an worker will take off the company (whittling down) or not based on various highlights given within the dataset.

Splitting Dataset:

Firstly, we part the dataset into two subsets, a preparing set and a testing set. The preparing set is utilized to prepare the machine learning models, whereas the testing set is utilized to assess their execution. Regularly, around 70-80% of the information is designated to the preparing set, and the remaining parcel is relegated to the testing set.

Model Selection:

For double classification assignments like foreseeing representative whittling down, a few machine learning calculations can be considered. In this examination, we select three commonly utilized calculations:

1. **Logistic Regression:** Logistic regression is a simple yet powerful algorithm for binary classification. It models the probability of a binary outcome using a logistic function, making it suitable for predicting binary outcomes like employee attrition.

2. **Random Forest Classifier:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It is known for its robustness and ability to handle complex datasets with high-dimensional feature spaces.

3. **Gradient Boosting Classifier:** Gradient boosting is another ensemble learning technique that builds a strong predictive model by sequentially adding weak learners (decision trees) to correct the errors made by previous models. It is particularly effective in improving model performance and reducing bias and variance.

By implementing logistic regression, random forest, and gradient boosting classifiers, we aim to build robust predictive models that can accurately predict employee attrition and provide valuable insights for businesses to mitigate attrition risks and retain their valuable talent.

Model Evaluation:

After training and evaluating the models, we obtained the following performance metrics:

1. **Logistic Regression:**
   - Accuracy: 83.20%
   - Precision (Class 0): 83%
   - Precision (Class 1): 83%
   - Recall (Class 0): 84%
   - Recall (Class 1): 83%
   - F1-score (Class 0): 83%
   - F1-score (Class 1): 83%

2. **Random Forest Classifier:**
   - Accuracy: 92.71%
   - Precision (Class 0): 89%
   - Precision (Class 1): 97%
   - Recall (Class 0): 97%
   - Recall (Class 1): 88%
   - F1-score (Class 0): 93%
   - F1-score (Class 1): 92%

3. **Gradient Boosting Classifier:**
   - Accuracy: 91.50%
   - Precision (Class 0): 89%
   - Precision (Class 1): 95%

- Recall (Class 0): 96%
- Recall (Class 1): 87%
- F1-score (Class 0): 92%
- F1-score (Class 1): 91%

Feature Importance:

Analysis of feature importance in the random forest model revealed the following key factors influencing employee attrition:

- **Job Satisfaction:** Employees with lower job satisfaction are more likely to leave the organization.
- **Work-Life Balance:** Poor work-life balance contributes to higher attrition rates.
- **Years of Experience:** Employees with less experience are prone to attrition compared to their more experienced counterparts.

Model Optimization

To optimize the model further, we explore techniques such as hyperparameter tuning and feature selection. By fine-tuning the model parameters and selecting relevant features, we aim to improve the model's performance and generalization capability.

Results

After training and evaluating the models, we find that the random forest classifier achieved the highest accuracy of 92.7%. This indicates that the model correctly predicted employee attrition status for approximately 92.7% of the observations in the test set. Furthermore, the precision, recall, and F1-score for both classes are also high, indicating a well-performing model.

Upon analysing the feature importance in the random forest model, we find that certain factors such as job satisfaction, work-life balance, and years of experience have a significant impact on employee attrition. Employees with lower job satisfaction and poor work-life balance are more likely to leave the organization, while those with more experience tend to stay.

Conclusion

In conclusion, predictive analytics can play a crucial role in identifying and addressing employee attrition in organizations. By analyzing factors contributing to attrition and developing predictive models, businesses can proactively identify employees at risk of leaving and implement strategies for retention.

The analysis of the IBM HR Analytics Employee Attrition & Performance dataset revealed valuable insights into the factors influencing attrition, including job satisfaction, work-life balance, and years of experience. By leveraging machine learning algorithms such as random forest, businesses can accurately predict employee attrition and take proactive measures to retain valuable talent.

Future Work

In future work, we can explore advanced machine learning techniques such as neural networks and ensemble methods to further improve predictive accuracy. Additionally, conducting longitudinal studies to track employee attrition over time and analysing trends can provide deeper insights into underlying causes and potential interventions.