IBM Innovation Center for Education

PHEMESOFT
Software That Matters
TRUSTED DELIVERY PARTNER for IBM ICE

Sri Sri University

Employee Attrition Prediction

PROJECT SYNOPSIS

BACHELOR OF TECHNOLOGY
Computer Science (AIML & DS)

SUBMITTED BY

1. Samriddhi Jain - FET/BCE/2021-25/039
2. Ayush Ashok Patil - FET/BCE/2021-25/044
3. Shrestha Agarwal - FET/BCE/2021-25/040
4. Akash Samantray - FET/BCE/2021-25/010

GUIDED BY

Sudhir Kumar Mohapatr

TITLE

1. Name –

   Samriddhi Jain - FET/BCE/2021-25/039

   Ayush Ashok Patil - FET/BCE/2021-25/044

   Shrestha Agarwal - FET/BCE/2021-25/040

   Akash Samantray - FET/BCE/2021-25/010

2. Present Official Address – Sri Sri University, Gate No - 2, Ward No - 3, Godisahi, Cuttack, Odisha-745006

3. E-mail –

   samriddhi.j2021btcseai@srisriuniversity.edu.in

   ayush.p2021btcseai@srisriuniversity.edu.in

   shrestha.abtcseds@srisriuniversity.edu.in

   akash.s2021btcseai@srisriuniversity.edu.in

4. Phone no -

   +91 9119133553

   +91 7875966960

   +91 9193911757

   +91 9337147914

5. Branch – B.Tech Computer Science (AIML & DS)

6. Batch – 2021-25

7. Proposed Topic – Employee Attrition Prediction

INDEX

Introduction

Employee attrition is a critical concern for organizations across various industries, as it impacts workforce stability, productivity, and overall business performance. In an era where retaining top talent is pivotal for success, predictive analytics and machine learning techniques offer invaluable tools for proactively managing and mitigating employee turnover. This project focuses on leveraging machine learning to develop a predictive model for employee attrition, providing organizations with actionable insights to enhance employee retention strategies.This project lies at the intersection of human resources, data science, and machine learning. It involves applying advanced analytics to personnel data, aiming to identify patterns and factors influencing employee attrition. The intersection of these fields enables the creation of a robust model that not only predicts attrition but also provides a deeper understanding of the underlying dynamics contributing to employee turnover. Technical terms used were:

Machine Learning: The project employs machine learning algorithms, a subset of artificial intelligence, to analyze historical data and make predictions about future employee attrition.

Predictive Analytics: The use of statistical algorithms and machine learning techniques to identify patterns and trends in data, enabling the prediction of future events, such as employee attrition.

Feature Engineering: The process of selecting, transforming, or creating relevant features (variables) from the dataset to enhance the predictive power of the machine learning model.

Model Evaluation Metrics: Metrics like Mean Squared Error (MSE), Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC) are used to assess the performance of the machine learning model.

Data Preprocessing: The cleaning and transformation of raw employee data, including handling missing values, encoding categorical variables, and ensuring data quality before feeding it into the machine learning model.

Decision Tree: A machine learning algorithm used for classification and regression tasks, visually represented as a tree-like structure where each node represents a decision based on input features.

ROC Curve: Receiver Operating Characteristic Curve illustrates the trade-off between true positive rate and false positive rate, helping to choose the optimal threshold for classification models.

Attrition Prediction Model: The final outcome of the project, a machine learning model capable of predicting the likelihood of employee attrition based on historical data and identified features.

By delving into the specialized field of employee attrition prediction, this project aims to empower organizations with actionable insights for enhancing employee retention strategies and fostering a more stable and engaged workforce.

Methodology

1. Data Acquisition:

The foundation of this project lies in the data utilized to train and evaluate the predictive model. Our primary data source will be IBM HR Analytics Employee Attrition & Performance. This data encompasses various aspects of employee life cycles, including: **Demographic information**: Age, gender, education, tenure, etc.**Performance metrics**: Individual performance evaluations, or productivity data. **Job satisfaction surveys**: Results from surveys capturing employee sentiments regarding work-life balance, compensation, career growth opportunities, and overall satisfaction.

2. Preprocessing and Feature Engineering:

Before delving into model building, the acquired data will undergo necessary preprocessing steps. Missing values might be handled through imputation techniques, outliers explored and potentially addressed depending on their nature and influence. Furthermore, data inconsistencies in format or coding will be rectified.

Feature Selection: Employing domain knowledge and techniques like correlation analysis, we will identify features likely to be associated with employee attrition. Additionally, feature importance scores from initial machine learning models can guide further refinement of the feature set.

3. Model Selection and Training:

Several machine learning algorithms will be explored for their effectiveness in predicting employee departures. Prominent candidates include:

Logistic Regression: A widely used algorithm for classification tasks, suitable for modeling the binary outcome of employee retention or attrition.

Decision Trees: Offering decision-making rules that are interpretable, these models can be insightful in understanding which factors most strongly influence employee decisions.

Random Forests: By combining multiple decision trees, these models aim to enhance robustness and reduce overfitting.

XGBoost: A tree-based boosting algorithm known for its high performance in various classification tasks, potentially leading to superior prediction accuracy.

The chosen models along with some additional models will be trained and evaluated using appropriate metrics like accuracy, precision, recall, and F1-score. Additionally, Area Under the ROC Curve (AUC-ROC) will be employed to assess the model's ability to distinguish between employees likely to leave and those who are likely to stay.To prevent overfitting, a commonly used technique is to split the data into training and validation sets. Typically, around 70% of the data will be used for training the model, while the remaining 30% serves as the validation set for model evaluation.

4. Evaluation and Analysis:

Evaluation goes deeper than just accuracy. We'll test the model on unseen data and analyze its decision-making process to understand which factors most influence its predictions. We'll also acknowledge potential biases and limitations to ensure our findings are reliable and relevant.

FACILITIES REQUIRED

Software Facilities:

1. Programming language: Python
2. IDE: Jupyter Notebook
3. Visualization Tools: Matplotlib, Seaborn
4. Machine Learning Tools: pandas, numpy, sklearn

Hardware Facilities:

1. RAM Capacity: Python programs can consume varying amounts of memory, particularly when dealing with large datasets or complex computations. Having sufficient RAM allows your system to handle the data efficiently, preventing slowdowns and improving the overall responsiveness of the development environment.
2. Sufficient Storage Capacity: Ample storage space is essential for storing Python scripts, libraries, datasets, and project files. Having enough storage capacity allows you to manage and organize your coding projects without worrying about running out of space.
3. High performance hardware:

- Central Processing Unit (CPU): A fast and multi-core CPU enhances the speed of executing Python code, especially for tasks involving data processing and computations.
- Graphics Processing Unit (GPU): While not essential for all Python programming, a powerful GPU can be beneficial for tasks related to machine learning, scientific computing, and data visualization.

References

1. Kaggle: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
2. Adaboost: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html
3. Bias-variance tradeoff: https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff
4. XG boost:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
5. SVM:https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
6. Random forest classifier:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
7. ROC:https://en.wikipedia.org/wiki/Receiver_operating_characteristic
8. Confusion Matrix:https://www.geeksforgeeks.org/confusion-matrix-machine-learning/