IBM Innovation Center for Education

PHEMESOFT
Software That Matters
TRUSTED DELIVERY PARTNER for IBM ICE

Sri Sri University

Project - High Level Design

on

Employee Attrition Prediction

Team members:

Ayush Patil
Shrestha Agarwal
Samriddhi Jain
Akash Samantray

Guided by: Mr. Sudhir Mohapatra

Industry Mentor:

Table of Contents

Scope of Document

The scope of this high-level design document encompasses a comprehensive analysis and prediction framework for employee attrition within an organizational context. This document aims to articulate the methodology, approach, and key components involved in conducting exploratory data analysis (EDA), statistical analysis, and the subsequent application of machine learning algorithms for predicting employee attrition. It will delineate the boundaries of the project, detailing the datasets used, analytical techniques employed, and the anticipated outcomes.

The document will primarily focus on elucidating the intricacies of the project's data-driven methodology, statistical models, and machine learning algorithms employed for predicting employee attrition. Additionally, it will provide insights into the underlying principles, data pre-processing strategies, and the interpretability of results. By defining the scope, the document ensures a clear understanding of the project's objectives and limitations, enabling a streamlined execution of subsequent phases.

Intended Audience

The intended audience for this high-level design document comprises a diverse group of stakeholders with varying levels of expertise in data science, analytics, and human resource management. This includes but is not limited to:

1. Data Scientists and Analysts: Professionals responsible for implementing the analytical techniques and machine learning algorithms. This audience requires a detailed understanding of the technical aspects of the project, including the choice of models, feature engineering, and evaluation metrics.

2. Human Resource Managers and Practitioners: Individuals involved in strategic decision-making related to workforce management. This audience seeks insights into the practical implications of the analysis, including potential interventions and proactive measures to mitigate attrition.

3. IT and System Administrators: Professionals responsible for implementing the technical infrastructure required for data storage, processing, and model deployment. This audience requires information on system requirements and integration considerations.

4. Project Managers and Executives: Individuals overseeing the project's progress and aligning it with organizational goals. This audience seeks a high-level understanding of the project scope, potential business impacts, and strategic implications.

System Overview

The system overview delineates the key components and interactions within the project. It serves as a roadmap for understanding the flow of activities, from data acquisition to model deployment.

The system comprises three main phases:

1. Exploratory Data Analysis (EDA): This phase involves a meticulous examination of the dataset to uncover patterns, trends, and potential influencing factors related to employee attrition. EDA provides a foundational understanding of the data, setting the stage for subsequent analytical endeavors.

2. Statistical Analysis: Building upon the insights gained from EDA, statistical analysis aims to quantify relationships, identify significant variables, and assess the impact of various factors on attrition. This phase involves the application of inferential statistical methods to derive meaningful conclusions.

3. Machine Learning Algorithm Implementation: The final phase involves the application of machine learning algorithms to predict employee attrition. This includes model training, validation, and evaluation. The document will detail the choice of algorithms, parameter tuning, and the rationale behind the selection.

Application Design

1. Data Exploration: Perform exploratory data analysis (EDA) on the provided dataset to understand the distribution of features, identify patterns, and gain insights into potential factors influencing employee attrition.

2. Data Preprocessing: Prepare the dataset for model training by handling missing values, encoding categorical variables, and scaling numerical features.

3. Feature Engineering: Extract relevant features and create new ones that might enhance the model's predictive capabilities.

4. Model Selection: Choose appropriate machine learning algorithms for building the attrition prediction model. Evaluate different models and select the one with the best performance.

5. Model Training: Train the selected model on the training dataset, using appropriate techniques such as cross-validation to optimize model parameters.

6. Model Evaluation: Evaluate the trained model on a separate test dataset to assess its performance in terms of accuracy, precision, recall, and other relevant metrics.

7. Prediction: Use the trained model to make predictions on new data, identifying employees at risk of attrition.

8. Documentation: Provide clear and concise documentation within the Jupyter Notebook, explaining the methodology, code rationale, and results.

Process Flow

Predicting employee attrition using machine learning involves several steps in the process. Below is a generalized process flow for building a machine learning model to predict employee attrition:

1. Problem Definition and Goal Setting:
   - Clearly define the problem you want to solve: predicting employee attrition.
   - Set specific goals for the prediction model, such as accuracy, precision, recall, or F1 score.
2. Data Collection:
   - Gather historical data related to employee turnover.
   - Include relevant features such as employee demographics, job satisfaction, performance metrics, and any other factors that may influence attrition.
3. Data Preprocessing:
   - Handle missing values: impute or remove incomplete data.
   - Encode categorical variables using techniques like one-hot encoding.

- Scale numerical features if needed.
- Explore and visualize the data to identify patterns and relationships.

4. Feature Engineering:
- Create new features that might enhance the predictive power of the model.
- Remove irrelevant or redundant features.
- Consider interactions between features.

5. Data Splitting:
- Split the dataset into training and testing sets. A common split is 80-20 or 70-30.

6. Model Selection:
- Choose a suitable machine learning algorithm for the task. Common models for binary classification tasks like attrition prediction include logistic regression, decision trees, random forests, support vector machines, and gradient boosting algorithms.

7. Model Training:
- Train the selected model on the training dataset.
- Use cross-validation techniques to assess the model's performance and tune hyperparameters.

8. Model Evaluation:
- Evaluate the trained model on the testing dataset.
- Use metrics such as accuracy, precision, recall, F1 score, and ROC-AUC to assess performance.

9. Model Interpretation:
- Understand the factors contributing to predictions.
- Explore feature importance to identify key variables.

10. Hyperparameter Tuning:
- Optimize the model's hyperparameters to improve performance.

11. Model Deployment:
- Deploy the trained model to a production environment for real-time predictions.
- Integrate the model into existing systems or workflows.

12. Monitoring and Maintenance:
- Regularly monitor the model's performance in the production environment.
- Retrain the model periodically with new data to ensure it stays accurate and relevant.

Information Flow

1. User Input:
- Stakeholders provide input regarding the requirements, objectives, and desired features of the attrition prediction system.

2. Data Collection and Integration:
- Collect historical data related to employee turnover.
- Integrate data from various sources, including HR databases, employee surveys, and performance records.

3. Data Preprocessing:
- Cleanse and preprocess the data to handle missing values, outliers, and inconsistencies.
- Conduct exploratory data analysis to understand the data distribution and relationships.

4. Feature Engineering:
- Create new features or transform existing ones to improve the model's predictive power.
- Consider factors such as employee demographics, job satisfaction, performance metrics, and tenure.

5. Data Storage:
- Store the preprocessed and engineered data in a suitable database or data warehouse for easy retrieval and analysis.

6. Model Development:
- Select appropriate machine learning algorithms (e.g., logistic regression, decision trees, random forests) based on the problem requirements.
- Develop models using the training dataset and perform cross-validation for model evaluation.

7. Model Training:

- Train the chosen machine learning model(s) using the preprocessed data.
- Fine-tune hyperparameters for optimal performance.

8. Model Evaluation:

- Assess the model's performance using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC on a validation dataset.

9. Model Deployment:

- Deploy the trained model to a production environment.
- Integrate the model into the organization's systems, allowing for real-time predictions.

10. Monitoring and Logging:

- Implement monitoring mechanisms to track the model's performance and behavior in the production environment.
- Set up logging for capturing relevant information about predictions and system activities.

11. Reporting and Visualization:

- Create reports and visualizations to communicate predictions and insights to stakeholders.

Data Design

1. Data Fields:

- Employee Information:
- Employee ID
- Age
- Gender
- Marital status
- Education level
- Department
- Job role
- Years at the company
- Monthly income
- Overtime
- Over 18
- Work hours per week
- Work-life balance
- Job satisfaction
- Base salary
- Bonus
- Stock options
- Distance from Home

2. Data Types:

Categorical Variables:

- Gender
- Marital status
- Education level
- Department
- Job role

Numerical Variables:

- Age
- Monthly income

- Overtime hours
- Performance ratings
- Work-life balance rating
- Date Variables:
-
- Date of hire
- Date of the last promotion
- Date of leaving (if applicable)

3. Data Preprocessing:

- Handling Missing Data
- Decide on strategies for dealing with missing values (e.g., imputation or removal).
- Scaling
- Scale numerical features if using algorithms sensitive to feature scales.
- Label Encoding
- Convert categorical variables into numerical format.

4. Data Splitting:

Training and Testing Sets: Split the dataset into training and testing sets to assess model performance.

5. Feature Engineering:

6. Data Versioning:

Record Changes:

Implement a system to track changes in the dataset, especially if you regularly update your attrition model.

7. Data Privacy and Security:

Anonymization:

Anonymize or pseudonymize sensitive information to protect employee privacy.

Access Controls:

Implement access controls to restrict who can view or manipulate the attrition dataset.

8. Continuous Monitoring:

Monitoring Changes:

Regularly monitor changes in the distribution of features and the attrition rate to identify potential shifts in the workforce.

9. Documentation:

Metadata:

Document the metadata, including the source of data, data collection methods, and any transformations applied.

Data Model

1. Define the Problem:

Clearly define the problem one wants to solve. In this case, it's predicting employee attrition.

2. Data Collection:

Gather relevant data for both employees who left (attrition cases) and those who stayed. Include a mix of demographic, performance, and job-related features. Sample features may include:

- Age
- Gender
- Job role
- Distance from home
- Years at company
- Environment Satisfaction
- Monthly income
- Hourly rate
- Percent salary hike
- Work-life balance

3. Data Preprocessing:

Clean and preprocess the data to handle missing values, outliers, and categorical variables. Convert categorical variables into numerical format using techniques like label encoding although our data set was clean without any missing values or noises.

4. Feature Engineering:

Create new features that might enhance the predictive power of the model.

5. Split the Data:

Split the dataset into training and testing sets to evaluate the model's performance on unseen data. We have split the dataset into training and testing in 70:30

6. Choose a Model:

Select a machine learning algorithm suitable for binary classification problems like Random Forests, XGBoost, AdaBoost, Logistic Regression etc.

7. Model Training:

Train the selected model on the training data. Use techniques like cross-validation to optimize hyperparameters and ensure robust performance.

8. Model Evaluation:

Evaluate the model using the testing set. Common metrics for binary classification include accuracy, precision, recall, F1 score, and ROC-AUC.

9. Interpretation and Insights:

Examine feature importance to understand which factors contribute most to attrition predictions. This can provide insights for HR strategies.

10. Deployment:

Deploy the model into production if it meets performance requirements. Integrate it into HR systems for ongoing attrition prediction.

11. Continuous Monitoring and Updating:

Keep monitoring the model's performance over time and update it as needed. Employee behaviors and company dynamics may change, requiring adjustments to the model.

Additional Considerations:

Ethical considerations: Ensure fairness and avoid biased predictions.

Explainability: Choose models that provide interpretable results, especially in HR-related decisions.

References

- Kaggle
- Sklearn documentation
- Medium
- Analyticsvidhya
- Geeksforgeeks
- Google Research papers