HR Analytics (Employee Attrition Analysis)

Low Level Design

Domain: Business Analytics

Creator: Md Al-Hasib Faisal

Date: 25/3/2024

Dhaka, Bangladesh

Email: hasibfaisal130@gmail.com

WhatsApp: +8801839586243

Document Version Control

Date issued	Version	Description	Author
1/3/2024	1.1	Initial Draft	Md Al-Hasib Faisal
3/3/2024	1.2	Modified final draft	Md Al-Hasib Faisal

Contents

Introduction

What is Low-Level Design Document?

The goal of LLD or a low-level design document is to give the internal logical of the actual program code for the HR Analytics (Employee Attrition Analysis). It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli.

The main objective of the project is to address inquiries such as:

- Patterns in employee turnover
- Time taken for employee recruitment
- Investment required for employee productivity
- Prediction of employee attrition
- Evaluation of learning and development program efficacy

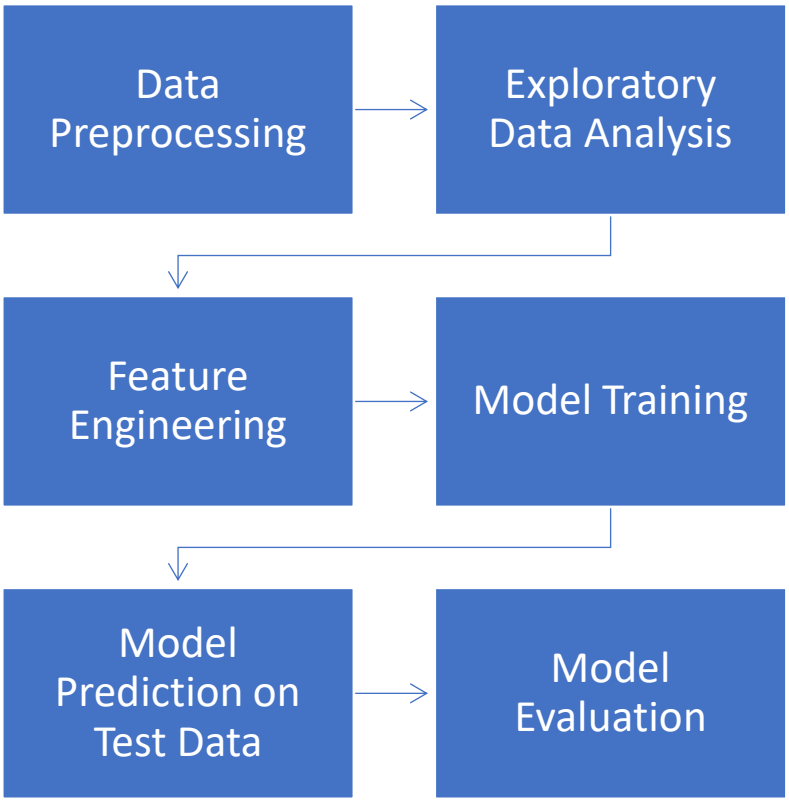
Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work.

Architecture

Data Preparation and Exploration

Model Development



Architecture Description

Data Preparation

Data Description

This dataset offers a comprehensive and varied analysis of an organization's employees, focusing on areas such as employee attrition, personal and job-related factors, and financials. Included are numerous parameters such as Age, Gender, Marital Status, Business Travel Frequency, Daily Rate of Pay, Departmental Information such as Distance From Home Office or Education Level Obtained by the employee in question. Also included is a variant series of parameters related to the job being performed such as Job Involvement (level), Job Level (relative to similar roles within the same organization), Job Role specifically meant for that individual(function/task), total working hours in a week/month/year be it overtime or standard hours for a given role. Furthermore detailed aspects include Percent Salary Hike during their tenure with the company from promotion or otherwise , Performance Rating based on specific criteria established by leadership , Relationship Satisfaction among peers at workplace but also taking into account outside family members that can influence stress levels in varying capacities ,Monthly Income considered at its starting point once hired then compared against their monthly payrate with overtime hours included if applicable along with Number Companies Worked before if any. Lastly the Retirement Status commonly known as Attrition is highlighted; covering whether there was an intent to stay with one employer through retirement age or if attrition took place for reasons beyond one’s control earlier than expected. Through this dataset you can get an insight into various major aspect regarding today's workforce management philosophies which have changed drastically over time due to advancements in technology.

Data Preprocessing

This module is responsible for loading the dataset, performing data preprocessing tasks, and preparing the data for further analysis and modeling.

Components:

- **Load Data**:
    - Reads the HR analytics dataset from a CSV file using pandas.
- **Remove Unnecessary Columns**:
    - Removes unnecessary columns ('EmployeeCount', 'EmployeeNumber', 'StandardHours') from the dataset.
- **Convert Attrition to Binary**:
    - Converts the 'Attrition' column values from 'Yes' and 'No' to binary values (1 and 0).

Exploratory Data Analysis

For EDA we did the following:

- Boxplot visualization of every numerical variable with respect to employee attrition.
- Density distribution of every variable with respect to employee attrition.
- Countplot of every variable w.r.t employee attrition.
- Horizontal barplot of every categorical variable against every numerical variables w.r.t employee attrition.

Feature Engineering

Train-Test Split Module

Description:

This module splits the preprocessed data into training and testing sets for model training and evaluation.

Components:

- **Split Data**:
    - Splits the preprocessed data into training and testing sets using **train_test_split** from sklearn.model_selection.

Column Transformer Module

This module performs feature engineering tasks, including identifying categorical and numerical columns, and preprocessing the data using a **ColumnTransformer**.

Components:

- **Identify Categorical and Numerical Columns**:
    - Uses pandas to identify categorical and numerical columns in the dataset.
- **Preprocessing with ColumnTransformer**:
    - Applies preprocessing transformations such as standard scaling for numerical features, one-hot encoding for categorical features, and LDA (Linear Discriminant Analysis to reduce dimensions) transformation.

Model Development

Model implementation

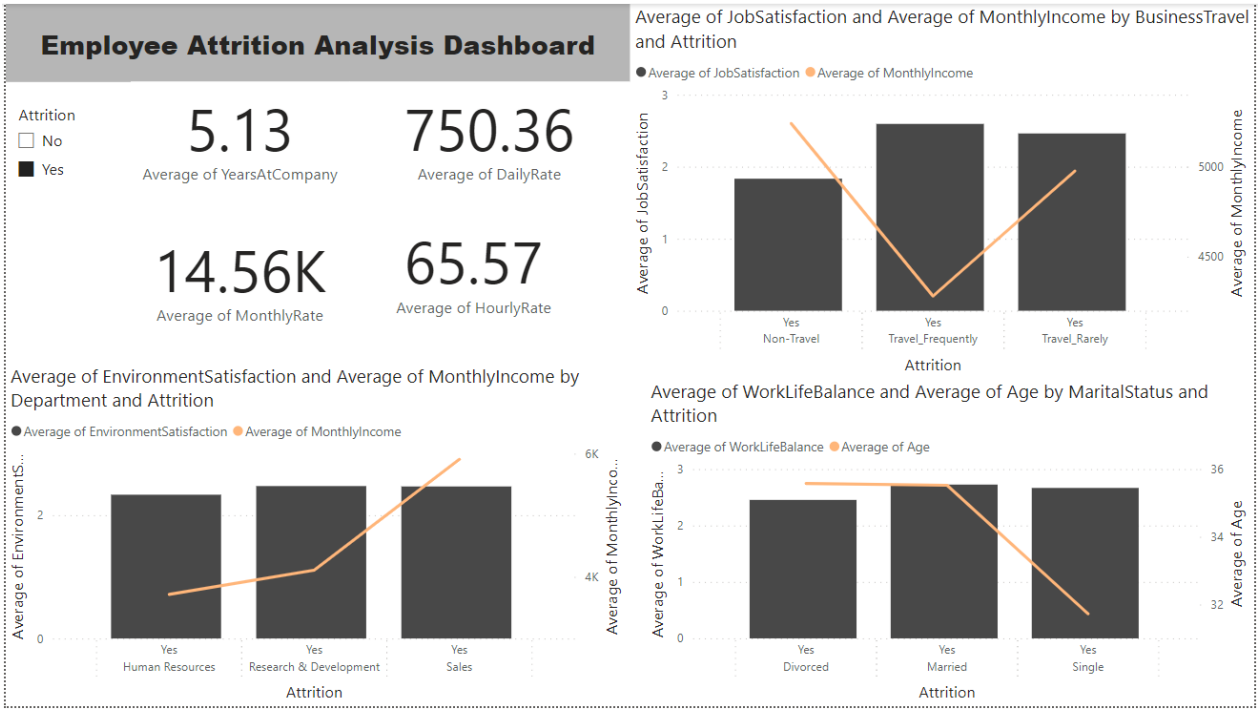
After column transformation is done, pipeline containing Standard Scaler and One Hot Encoder was fitted to several models such as Logistic Regressor, XGBoost, AdaBoost, Light GBM, RandomForest, Support Vector Machine and Stochastic Gradient Descent Classifier Models. They were all hyperparameter Tuned and we evaluated the classification model metrics by focusing on the Precision, Recall and F-1 score. Since we didn't want to lose any employees, we focused on Recall value as this is given more priority when we want less type-2 errors. After evaluating all the models, we've seen that Stochastic Gradient Descent Classifier model (without the hyperparameter tuning) works better than all the models.

Model Evaluation

Test dataset is used to evaluate the model. 20% of dataset was separated for testing. Predicted results of the model are compared with the actual data to check the amount of error. For the Stochastic Gradient Descent Classifier model we got Precision = 0.67, Recall = 0.75 and F-1 score

= 0.70. This is the best model so far. As the dataset was highly imbalanced it was hard to get good predictions with other models. More data needs to be collected to create a more balanced dataset and get even better models.

Power BI Dashboard:



Unit cases

Test Case Description	Expected Result
Employee attrition slicer (page-1)	When clicked (options : yes or no) it should show the bar charts having average job+environment satisfaction and work life balance w.r.t different categorical variables such as frequency of business travels, marital status and job department. It will also a line plot indicating the average monthly income.
Employee attrition slicer (page-1)	When clicked (options: yes or no) it should show the average years at the company, daily+hourly+monthly rate of payment
Employee attrition slicer (page-2)	When clicked (options: yes or no) it should show the average of years since last promotion,age,total working years against the different job roles, gender and education field w.r.t employee attrition. Also the line plots will indicate average daily rate, hourly rate and monthly rate.
Employee attrition slicer (page-3)	When clicked (options: yes or no) it should show the average distance from home, job and environment satisfaction against different job roles and if the employees work overtime or not w.r.t employee attrition. The line plots indicate average training times per year for each job role and average daily rate for working overtime.

Employee attrition slicer (page-4)	When clicked (options: yes or no) it should show the area plots of average years since last promotion against various job levels and years with current manager w.r.t the employee attrition.