Absenteeism Prediction with Logistic Regression

Category: Machine Learning, Classification

Date: 22nd March 2024

Introduction:

Absenteeism in the workplace can have significant implications for productivity, morale, and overall organizational performance. Detecting patterns of absenteeism early can enable proactive interventions and resource allocation to mitigate its impact. In this project, I propose to utilize logistic regression, a robust machine learning algorithm, to develop a predictive model for absenteeism in the workplace.

Logistic regression is a powerful tool for binary classification problems, where the objective is to predict one of two possible outcomes. In the case of absenteeism prediction, I train our logistic regression model on a dataset containing various attributes such as Age, Body Mass Index (BMI), Transportation Expense, and Daily Work Load Average, among others. These attributes will serve as input features, and our model will learn to distinguish between moderate and excessive absenteeism based on these characteristics.

Phase One: Data Preparation

In the initial phase of the project, our focus will be on gathering and preparing the requisite data for training and evaluating the logistic regression model for absenteeism prediction. I collected data from relevant sources, ensuring it is in a suitable format and ready for training the logistic regression model.

Phase Two: Training a Logistic Regression Model for Absenteeism Prediction

In the second phase, I will leverage the prepared data to train a logistic regression model capable of predicting the probability of moderate or excessive absenteeism based on employee attributes. This phase will encompass the following key steps:

Data Splitting: The dataset will be divided into two subsets: a training set and a testing set. The training set will facilitate model training, while the testing set will be used to assess the model's performance.

Model Training: I applied the logistic regression algorithm to the training data. During this stage, the model will learn the relationships between input features (e.g., transportation expense, distance to work, age, daily workload average, body mass index, education, children, pets) and the target variable (moderate or excessive absenteeism). The goal is to estimate coefficients for each feature, enabling the calculation of the probability of absenteeism.

Model Evaluation: The performance of the trained logistic regression model will be evaluated using the testing set. I myself build a predict function according to accuracy for binary classification models, to assess the model's ability to predict absenteeism based on employee attributes.

By completing Phase Two, I will have developed a logistic regression model capable of predicting the probability of moderate or excessive absenteeism based on employee characteristics.

Phase Three: Developing and Deploying an Absenteeism_Module
In this final phase, I create an Absenteeism_Module housing the logistic regression model. This module will offer methods to load and clean raw data, along with a prediction method for absenteeism based on employee attributes. By streamlining deployment and ensuring easy integration, this module will empower organizations to proactively address absenteeism and enhance workforce management efficiency.

Conclusion:
Through this project, I aim to leverage logistic regression as a powerful tool for absenteeism prediction in the workplace. By accurately identifying employees at risk of moderate or excessive absenteeism, organizations can implement targeted interventions to mitigate absenteeism's impact and enhance overall productivity and employee well-being.

Thank for reading