Job Retention

1. Abstract

Employees are an essential component of any company and there are many people who leave their jobs. The job retention project aims to analyze the employees' information to build machine learning models. The models will predict whether the employees will leave the company or not. On the other hand, the analysis will help the company to define the features that impact job retention. The used dataset to build the model is called "HR_comma_sep" from the Kaggle platform. Finally, the models were evaluated using accuracy metrics. The percentages are 78.77% and 94.59%, respectively.

2. Design the solution

Kaggle platform provided a dataset called " HR_comma_sep" that represents information from a particular company about their employees. The project aims to analyze this information and build a model to help the company to predict which employee will leave the company. Also, it will determine the variables that impact job retention.

3. Dataset

The dataset consists of 14999 data points with 10 features, two of which are categorical. The features of the dataset are satisfactory level, the number of projects, last evaluation, average monthly hours, time spend company, work accident, left, promotion last 5 years, department, and salary. After conducting an analysis of these features, it was found that eight of them are significant to use on the models.

4. Algorithm

4.1 Feature Engineering

1) Eliminate nonimportant features
2) Convert categorical variables into binary dummy variables.

4.2 Models

There are two models were used to make the prediction which are Logistic regression and k-nearest neighbors. However, the Logistic regression model was used based on the recommendation on Kaggle.

4.3 Model Evaluation and Selection

Splitting the dataset into train and test using the train test split library. The training size is 75% and the test size is 25%. The metric used to evaluate the model performance is classification report because the data is imbalanced. For the Logistic regression model, the result of the recall and f1-score are 85% and 81%, respectively. Whereas the k-nearest neighbors' model, the result of the recall and f1-score are 99% and 98%, respectively.

5. Tools

The following are tools that used in order to build the model:

- Numpy and Pandas for data manipulation.
- Scikit-learn for modeling.
- Matplotlib and seaborn for plotting.