HR Promotion Prediction

Abstract

This project amid to analyse the data related to employees of a massive organization to predict which employees will be promoted. Naturally, the variable is unbalanced: we have few promotions compared to employees who are not promoted, as is to be expected in any organization.

Kaggle provides the used data in this project, and the data will be compared to the scores of multiple ranking algorithms. With the sklearn library, the random forest was trained and got 93% accuracy.

Design

This project is one of the T5 Data Science BootCamp- T5o20- requirements. Data provided by Kaggle has been used in this project.

Data

The dataset is provided in .csv format. This dataset contains about +23,400 employees' records from a large multinational corporation (MNC), and they have 9 broad verticals across the organisation. and 13 features (columns). This dataset contains the information of each employee about:

1. employee_id Unique ID for employee
2. department Department of employee
3. region Region of employment (unordered)
4. education Education Level
5. gender Gender of Employee
6. recruitment_channel Channel of recruitment for employee
7. no*of*trainings no of other trainings completed in previous year on soft skills, technical skills etc.
8. age Age of Employee
9. previous*year*rating Employee Rating for the previous year
10. length*of*service Length of service in years
11. KPIs_met >80% if Percent of KPIs(Key performance Indicators) >80% then 1 else 0
12. awards_won? if awards won during previous year then 1 else 0
13. avg*training*score Average score in current training evaluations
14. is_promoted (Target) Recommended for promotion

Algorithms

Feature Engineering

- Explore the features
- Drop any unnecessary features
- Clean the data
- Data Analysis and Visualization

Models

- Data cleansing to train the model
- Convert the categories variables into Numerical variables
- Train and compare the scores of multiple ranking algorithms (Logistic Regression, Linear SVC, Decision Tree Classifier, Random Forest Classifier, Naive Bayes Classifier, K-nearest Neighbors Classifier)
- Apply "Random Forest Classifier".
- . The official metric was the accuracy of the model, where the model tested on the accuracy, precision, recall, and F1 score. The result of used model:
    - Accuracy: 93%
    - Precision: 75%
    - Recall: 32%
    - F1: 45%

Tools

- Numpy
- Pandas for data manipulation
- Scikit-learn for modeling
- re for clean data
- Matplotlib for plotting
- streamlit for interactive visualization