Predicting qualified employees for promotion using Classification models

Banan Alhethlool
banan.alhethlool@gmail.com

Randa Mohammed
randa1414@gmail.com

Abstract:

The purpose of this project is to use a large dataset and apply classifications algorithms to it. This project uses the HR of multinational corporation dataset of employees in order to predict the eligible employees for promotion. Hence save time and effort and expedite the process of promotions in the company. The prediction of promoted employees will be done through the use of many machine learning algorithms.

Design:

The classification project progress will follow this way. First, the dataset that is used in this project is about employees in a multinational corporation. It is downloaded from Kaggle.com. Then, the data is explored doing exploratory data analysis in order to see the relationship between different features and the target feature. The target feature in this project is to predict promoted employees from the dataset. The dataset is prepared by cleaning, feature engineering, and feature selection in order to use it in different classification algorithms.

Dataset: The data that will be used in this project is downloaded from Kaggle.com (https://www.kaggle.com/arashnic/hr-ana). The data is HR analytics data based on a multinational corporation with many departments. The dataset consists of over 50000 observations with 13 features.

Algorithms:

Data preparation

1. **Feature Selection** → Drop some columns which are recruitment_channel, region, gender and employee_id because they don't affect the prediction of the promoted employees.
2. **Feature Engineering** → Encoding & Scaling

   Encoding → We created dummy variables for education and departments columns since they have many categorical variables.

   Scaling → Standardization technique is used in order to make the features normal distribution.

3. **Balance Data** → SMOTE is used in order to oversample the minority class and make the data balanced.

Classification Models:

We have explored many models in order to find the best model that best suits our dataset. We have also tried many techniques in order to enhance the performance of the model. The models we have tried are:

- Logistic regression with AUC = 80.252
- KNN with AUC = 71.368
  - KNN grid search cv for best parameter
- Decision Tree with AUC = 64.147
- Random Forest with AUC = 77.498
- Adaboosting with AUC = 79.239
- Gradient Boosting with AUC = 79.798
- Naive Bayes (GaussianNB) with AUC = 71.032
- Stacking different models (Logistic Regression, KNN, Random Forest, Extra trees)

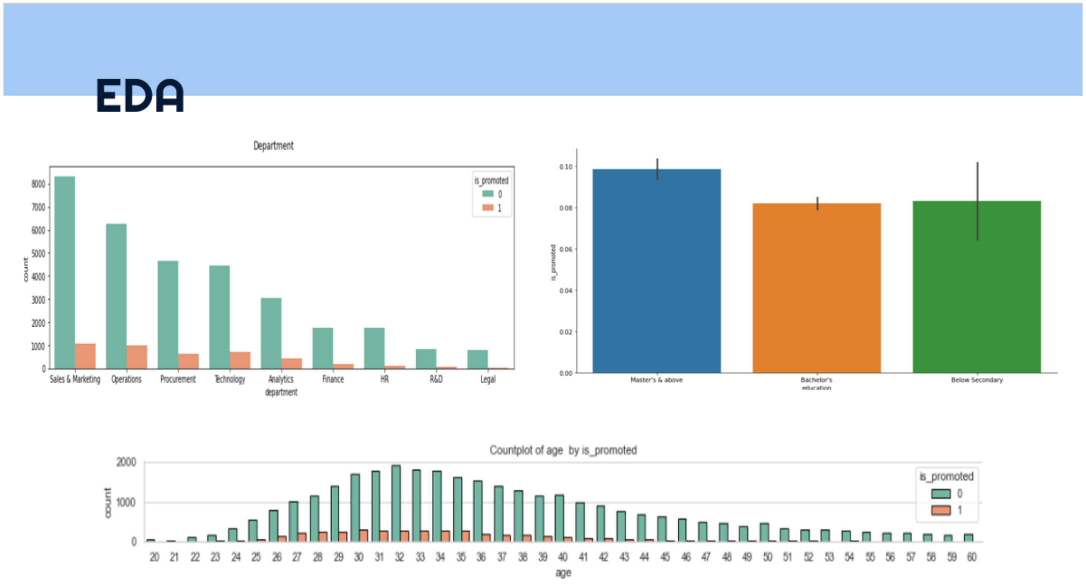
We did different techniques and applied many algorithms to the dataset in order to experiment and see the best model for our case. Comparing all models performed on the dataset. The different models were compared with the AUC metric, the best model is logistic regression because it has the higher AUC value which is equal to 80.252. The logistic regression is very simple and easy to interpret on data. It has the best performance in our case because the different values in the target class in the dataset can be separated by a line. The lowest AUC value belongs to the decision tree algorithm which is equal to 64.147. The decision tree can be improved by implementing different techniques such as using different hyper parameters such as depth of the tree in order to enhance its performance. Increasing the dataset size will improve the performance of the models. Since the model can learn more about the target from a larger dataset. All in all, the promotion is affected by a combination of factors, but the most important two features affecting the promotion is the performance of employees measured by awards granted. The second one is the number of training each employee gets.

Tools:

Technologies: Python, Jupyter Notebook.

Libraries: Pandas, Numpy, os, pickle, sklearn, imblearn.over_sampling, seaborn and matplotlib

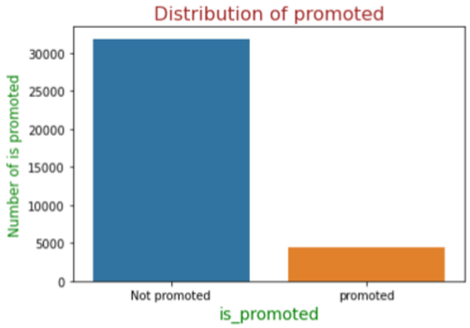
Communication:

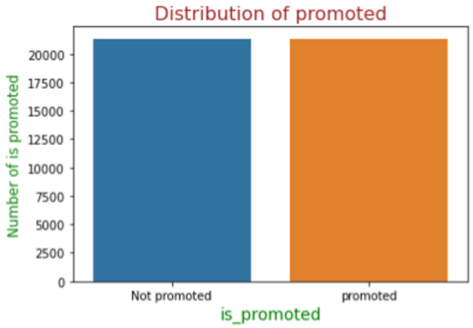


EDA

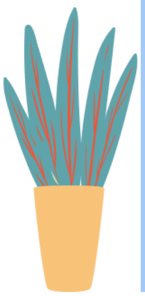
Balance Data

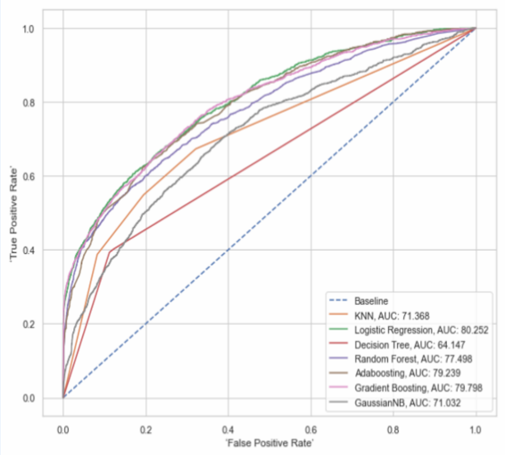
✓ **SMOTE**





Analysis & Results





Best Model Logistic Regression