Predicting Job Stability Using Classification

Leena AlQasem
Leenabdulh@gmail.com

Modhi AlHbrdi
modhi.alhbrdi@outlook.com

Abstract:

The third project of the Data Science Bootcamp T5 is called Predicting Job Stability Using Classification. Through the project, we will establish employee stability for SDAIA authority. Hence, this project aims to find the best ML models to predict employee stability based on many features.

Design:

As we are all heading to data science careers, our project called Predicting Job Stability Using Classification for the HR department in SDAIA aims to find the best ML models. Therefore, providing consistent, stable jobs can help companies increase retention, attract top talent, and create a positive work environment that brings out employees' best performance. Also, Save the efforts of hiring principles and do not lose their proficient employees

Data:

The datasets that was used in this project is a public source dataset, from Kaggle. Source: Here. The dataset includes many features it consists of *relevent_experience, company_size, training_hours, education_level, etc*. It is useful for the classification prediction.

- **Size:** 19159 records × 14 columns

Algorithms:

- ***Feature Selection:***
  - Drops both the 'enrolled_id' and the 'city' columns
- ***Feature Engineering:***
  - Encoding the columns into categorical values
  - Scaling using standard Scaler
- ***Balanced dataset:***
  - SMOTE was used for handling the imbalanced data

- ***Models:***

  For the result and analysis part, we did six models with cross-validations, and we obtained the four scores: accuracy, F-1, precision, and recall, as shown in table 1. We can tell that both the Random Forest and Gradient Boosting have the highest scores compared with other models.

Table 1 The models with their scores

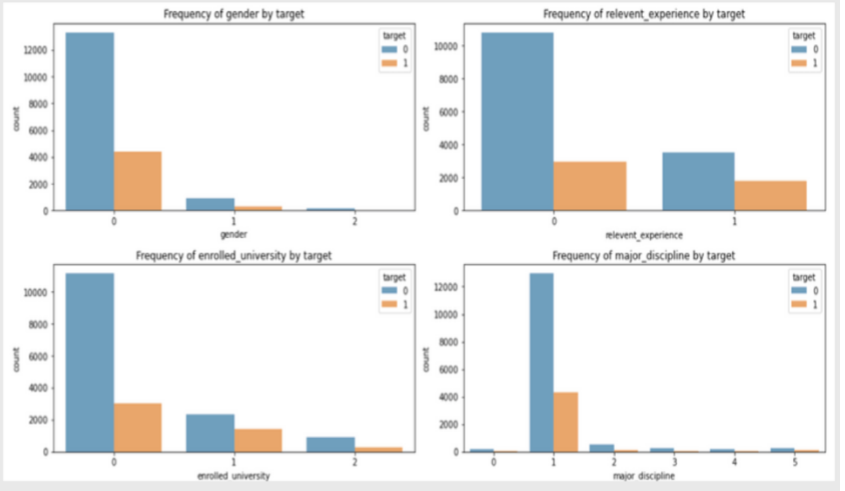
	Accuracy	F-1	Precision	Recall
KNN	0.79509	0.80914	0.75720	0.86873
Logistic Regression	0.68036	0.66348	0.70046	0.63021
Decision Tree	0.77064	0.7717	0.76799	0.77560
Random Forest	0.8350	0.83407	0.83887	0.82932
Gaussian Naive Bayes	0.66350	0.67434	0.65329	0.69679
Gradient Boosting	0.83617	0.83209	0.85337	0.81184

Tools:

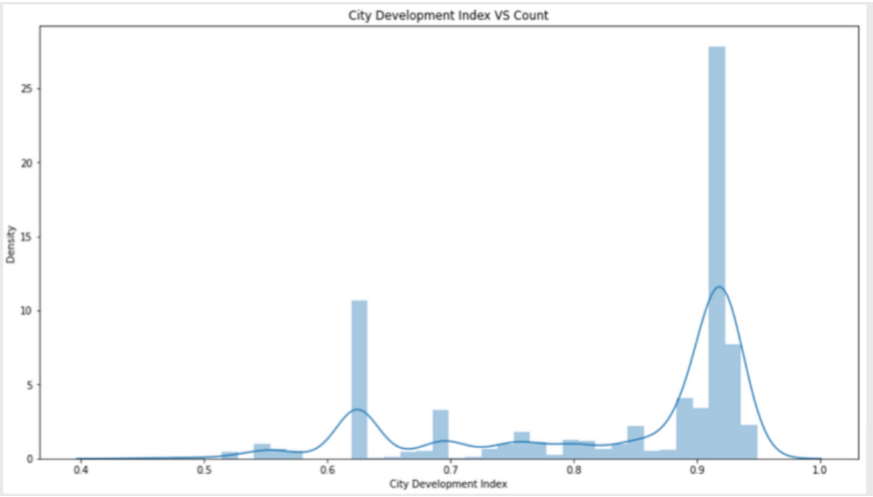
- **Technologies:** Jupyter Notebook, Python.
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn.

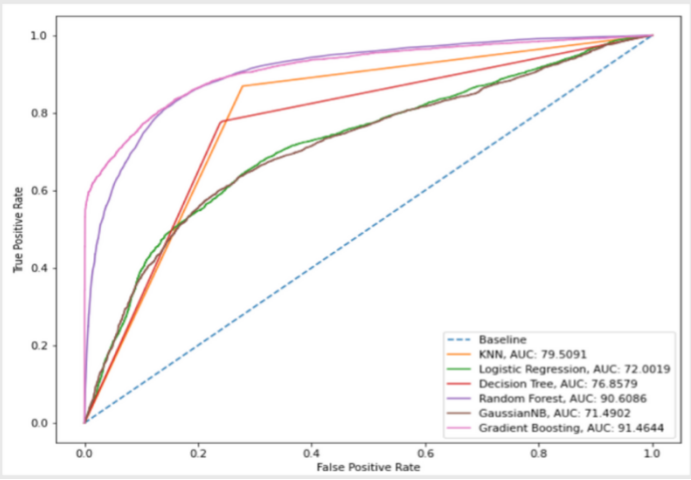
Communications:

A high proportion of males with related experience who have not attended college and are majoring in STEM are more likely to remain in the company.

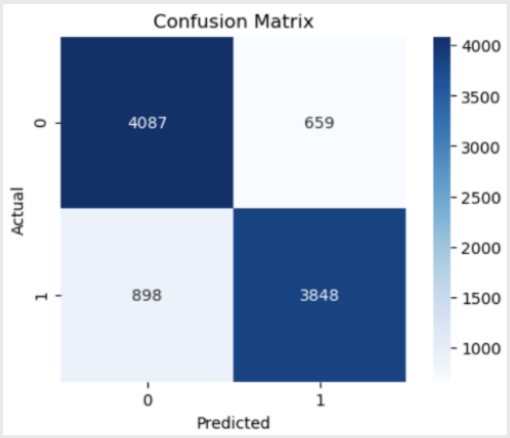


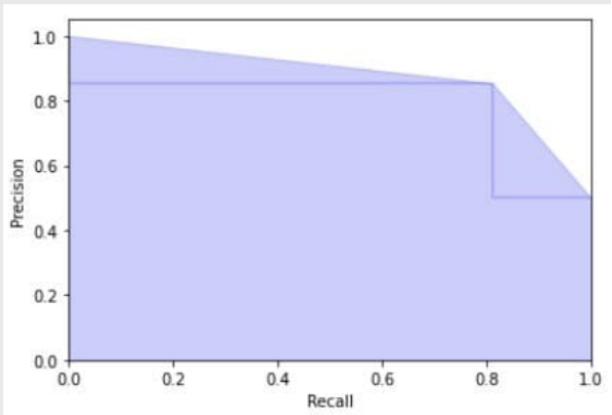
Employees are more likely to look for a new job in cities with a low development index and vice versa.





Gradient Boosting showed the highest AUC score with 91%. Followed by the Random forest with 90.6%





The Gradient Boosting Classifier predicts around 4000 not leaving the job correctly and around 3800 leaving the job correctly.

The Gradient Boosting Classifier with a higher AUC on the ROC curve will always have a higher AUC on the precision and recall curve.