SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية طويق
TUWAIQ ACADEMY

KAGGLE COMPETITION

Training Programs Dataset

Mohammed Alageel
Mazen Alamri

01. DATA DESCRIPTION

02. DATA EXPLORATION

03. DATA PREPROCESSING

04. MODEL BUILDING

05. MODEL IMPROVEMENT

TABLE OF CONTENT

DATA DESCRIPTION

The dataset is for trainees applying for training programs, it consists of 23 Features such as Age, Gender, Level of Education, etc.
and it has one target variable: whether the trainee completes the program or not.

DATA EXPLORATION

- The dataset contained 3 files: Train, Test for training and testing, and Registration file for student marks inthe registration process.
- After merging the 3 files, the resulting dataset contained 7366 rows and 37 columns.
- The dataset had many null values, more on the next slide

03.

DATA PREPROCESSING

FILLING NULL VALUES

- The dataset had null values in many features, including categorical and numeric ones.
- We dealt with null categorical values by filling them with the mode, and numerical values by filling them with the median.
- Some columns had a very high number null values, so we had to drop the columns.
- There were some columns that had null values, but we thought filling them with the mean or median was not correct, which we imputed using the KNN Imputer.

FEATURE ENGINEERING

As for feature engineering we added many features including:

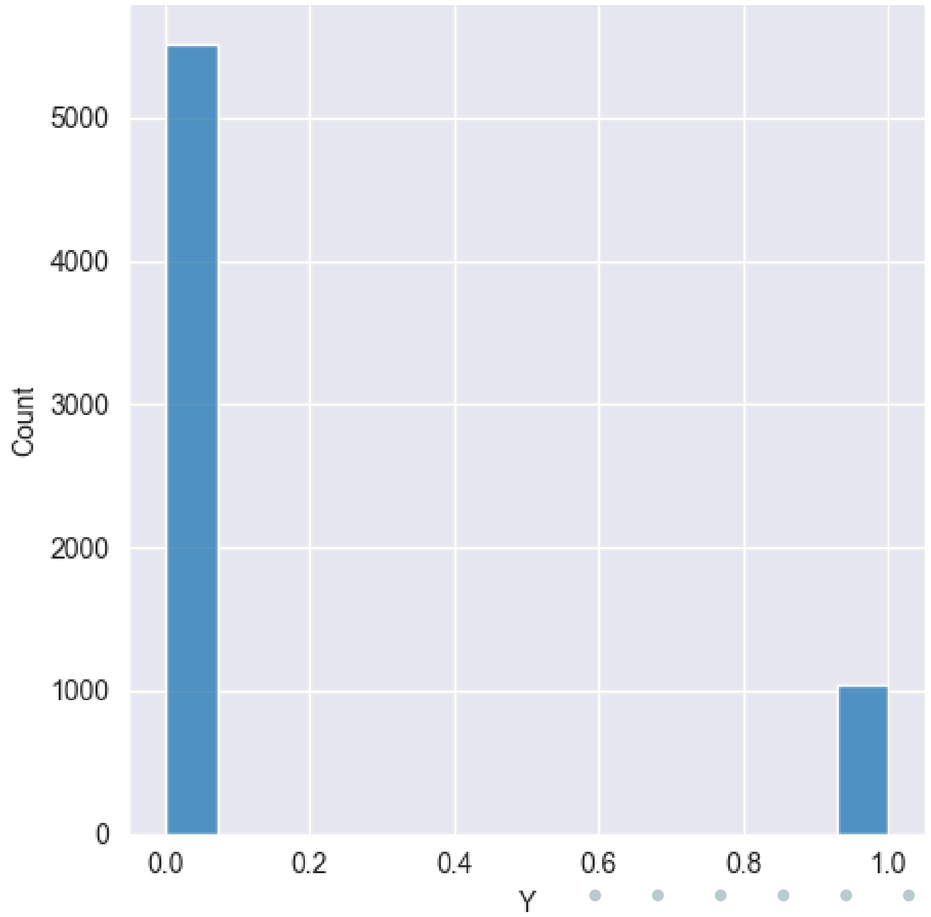
1. Program Start Year, Month, Day, etc.
2. Program End Year, Month, Day, etc.
3. The Program’s Duartion in months.
4. The Weighted University Score.
5. Whether the trainee is a Student.
6. Whether the trainee is an Employee.

ENCODING & NORMALIZATION

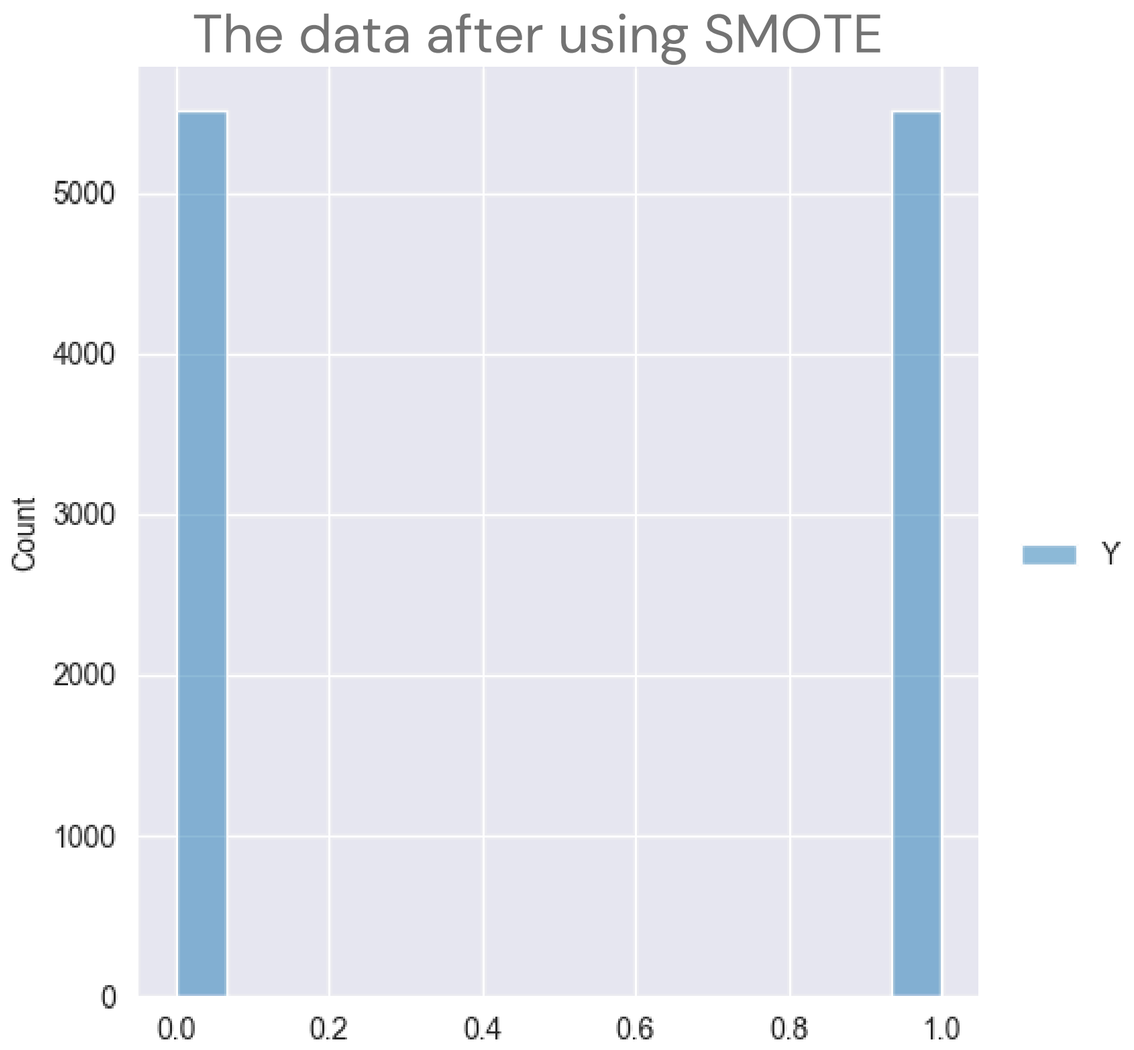
- We encoded variables like the program's skill level and level of education based on their values e.g. Higher education means a higher value and vice versa.
- As for other values we encoded them using an ordinal encoder.
- Normalized numeric features using a standard scaler.

DATA IMBALANCE

- We noticed that the data was imbalanced, so we used SMOTE (Synthetic Minority Oversampling Technique) to resample the data.
- We will discuss how it affected the models in a later slide.



DATA IMBALANCE



04.

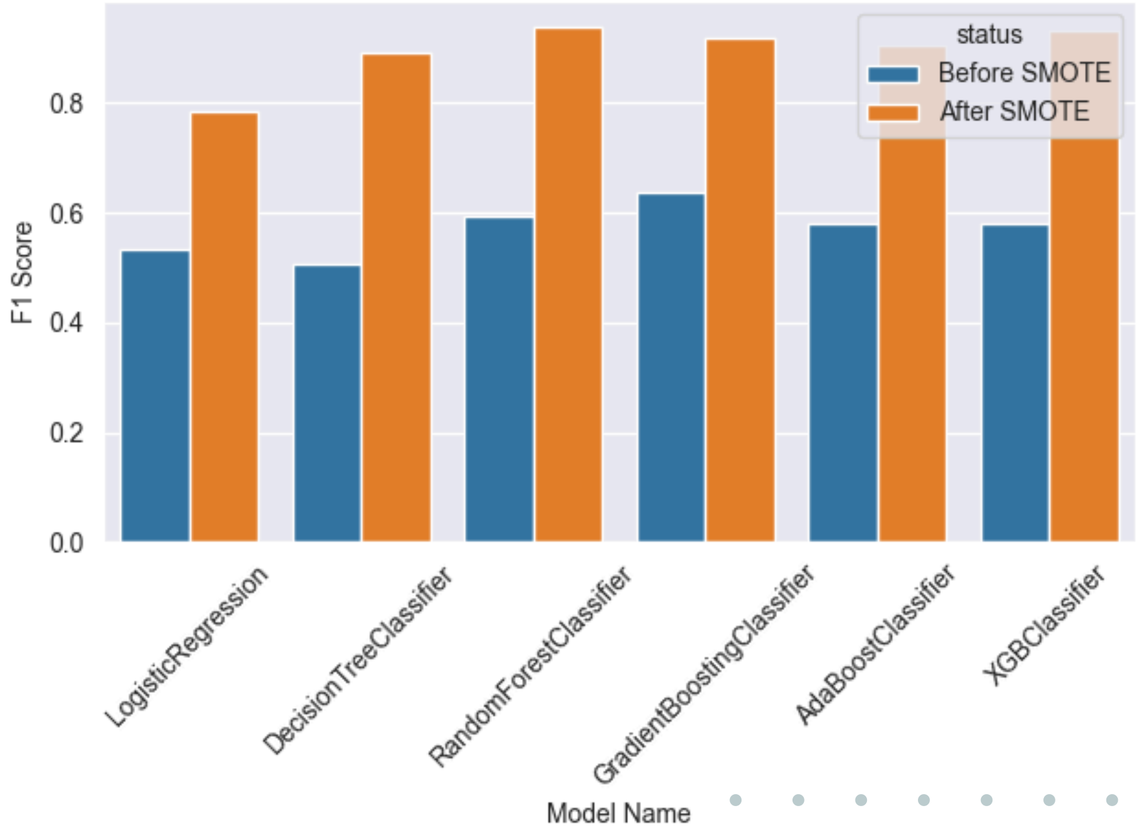
MODEL BUILDING

MODEL BUILDING & IMPROVEMENT

- We chose F1 score as an evaluation method for the classifier.
- We tried Logistic Regression and Decision tree models to see how the basic models preform.
- Then we tried other advanced models such as Random Forest, Gradient Boosting, Ada Boost, XGBoost.

DATA IMBALANCE

The following plot shows how the affect of data imbalance on the model's performance was big.



05.

MODEL IMPROVEMENT

HYPERPARAMETER TUNING

- The model with default hyperparameters had F1 score = 93.8%.
- Preformed hyperparameter on random forest using random search.
- With the tuned model on the hyperparameters we got from the random search it reached an F1 score of 94.3%.

THANK YOU