2021 HR Analytics Case Study

How “Anonymous” get a better prediction towards their candidates

Purwadhika Jakarta Batch 11

William Andreas

JOB CONNECTOR

Data Science

&

Machine Learning

The problem

Company

An anonymous company that's dynamic in Big Data and Data Science.

Context

- Needs to enlist data scientists.
- Needs to know which of these candidates are really wants to work for the company or not.

Problem statement

- To predict whether or not candidates alter their occupations after they have completed their training.
- Sorting out candidates that are false predicted

Challenges deep-dive

Challenge 1

Challenge 2

Challenge 3

Exploratory Data Analysis

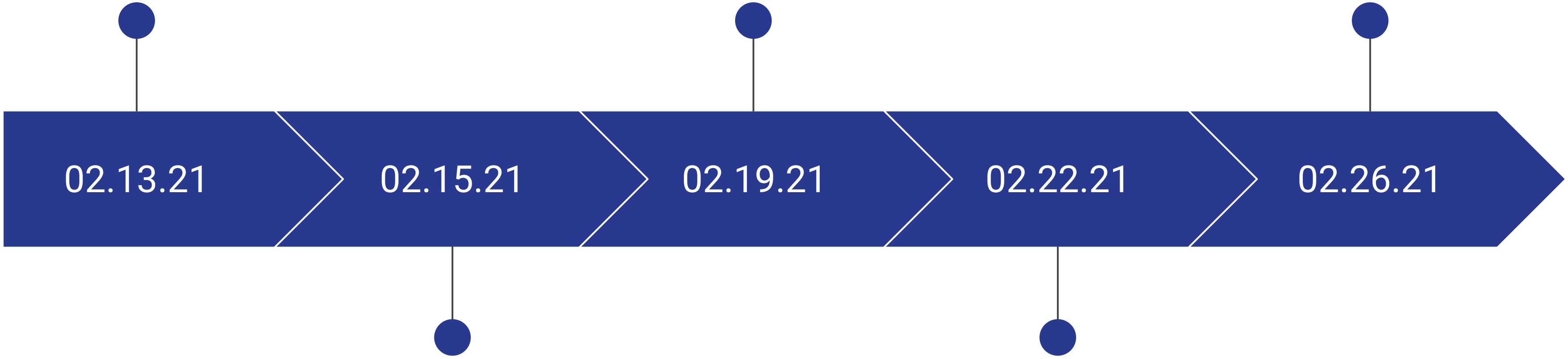
A thorough examination meant to uncover the underlying structure of a data set and exposes trends, patterns, and relationships that are not readily apparent.

Cleaning and Pre-Processing

- Transform the raw dataset into an understandable format.
- Improve data efficiency.

Modeling

Best possible predictive machine learning model to answer those in need.



02.13.21 Start final project

02.15.21 Exploratory Data Analysis

02.19.21 Data Cleaning and Preprocessing

02.22.21 Modeling and Dashboard

02.26.21 Due Date 16.00 PM WIB GMT+7

Tackle Challenges

Steps

1. Load Dataset
2. EDA
3. Handling Columns
4. Splitting Dataset
5. Handling Missing Values
6. Handling Outliers
7. Handling Imbalanced Data
8. Encoding
9. Feature Selection
10. Building Machine Learning Models
11. Hyperparameter Tuning
12. Evaluating Best Model
13. Saving and Deploy Model

Exploratory Data Analysis

Dataset

https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

enrollee_id	city
city_ development _index	gender
relevent_experience	enrolled_university
education_level	major_discipline
experience	company_size
company_type	lastnewjob
training_hours	target

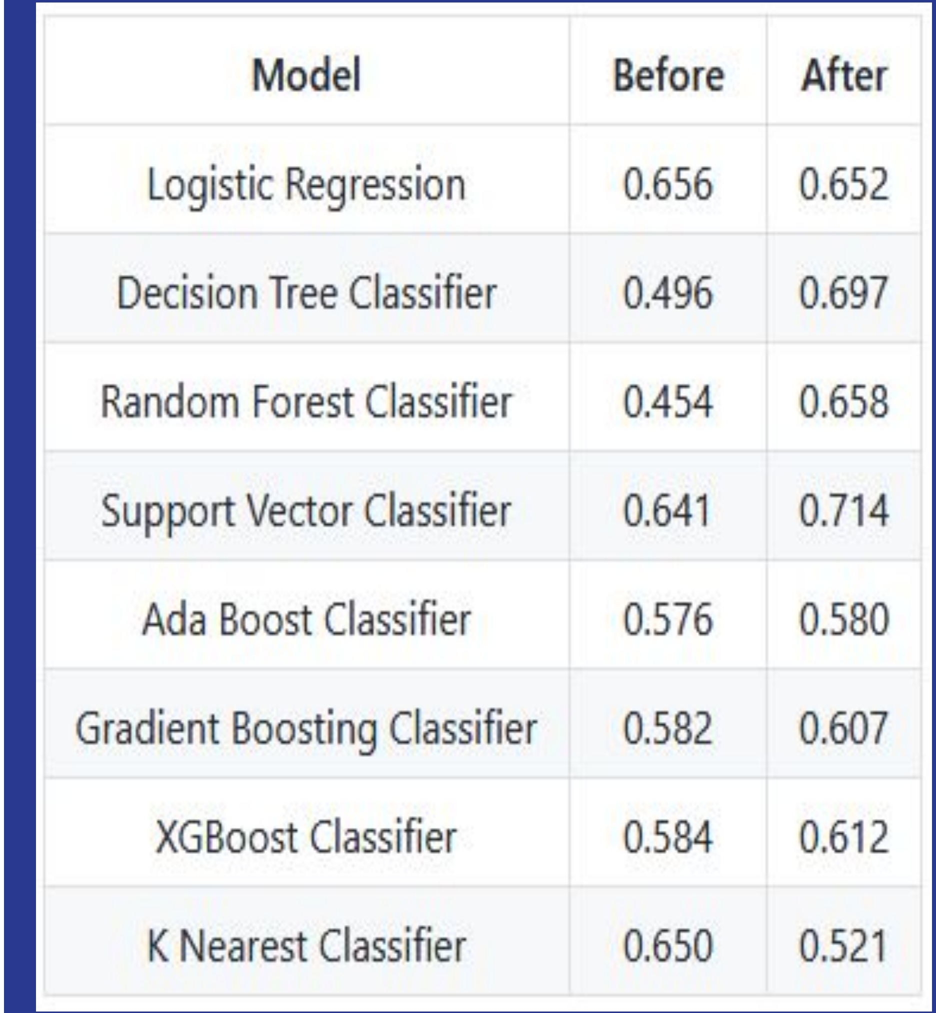
Cleaning and Pre-Processing

Steps

1. Drop rows containing ID and training hours of candidates.
2. Impute missing values using their most frequent values.
3. Encode categorical columns of data to numerical using Binary Encoder, and ordinal columns of data to numerical using Ordinal Encoder.
4. Scale numerical column of data using RobustScaler.
5. Generate polynomial and interaction features using PolynomialFeatures for numerical column of data after it's scaled.

Modeling

Supervised Machine Learning
Classification

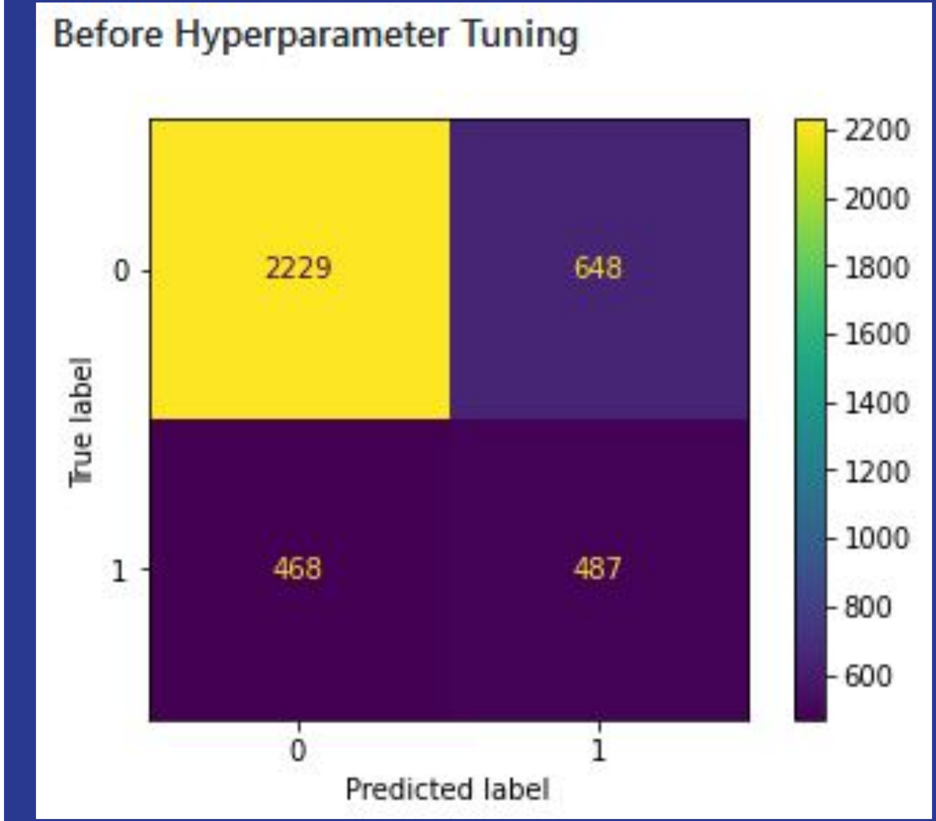


Model	Before	After
Logistic Regression	0.656	0.652
Decision Tree Classifier	0.496	0.697
Random Forest Classifier	0.454	0.658
Support Vector Classifier	0.641	0.714
Ada Boost Classifier	0.576	0.580
Gradient Boosting Classifier	0.582	0.607
XGBoost Classifier	0.584	0.612
K Nearest Classifier	0.650	0.521

Implementation

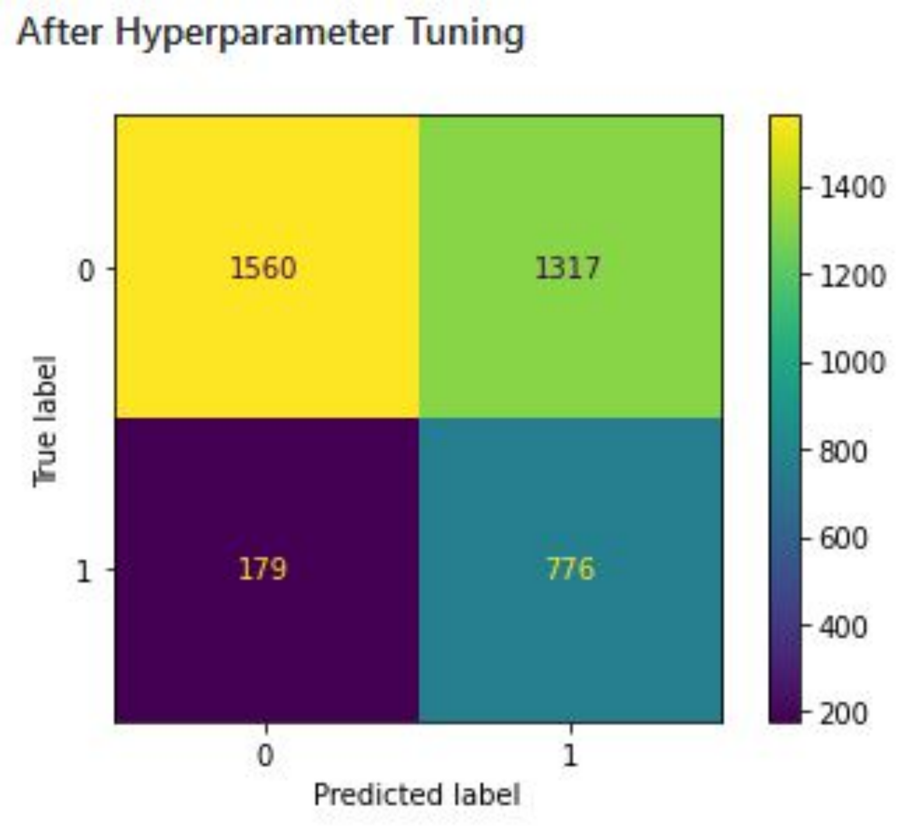
Decision Tree Classifier

Confusion Matrix



Decision Tree Classifier

Confusion Matrix



Decision Tree Classifier

Comparison

Predicted not changed actually changed

Before : 468/(2229+468)*100 = 17.35%

After : 179/(1560+179)*100 = 10.29%

Predicted changed actually not

Before : 648/(487+648)*100 = 57.09%

After : 1317/(776+1317)*100 = 62.92%

Impact

Trade-off
7.06% reduction in predicted not changed actually changed

5.83% addition in predicted changed actually not

Problem Solved!

Thank You