Predicting Job Stability For Data Scientist

Using Classification

Leena AlQasem *Modhi Alhbrdi*

Introduction

- *Brief Description*

  - ***Classification*** model to predict job stability of data scientist in SDAIA

  - *SDAIA became the core mandate to drive and own the national data and AI.*

- *Project Objectives:*

  - *Stable jobs, steady pay, and benefits.*
  - *Saves the efforts of the HR department.*
  - *Prevents losing proficient employees.*

Methodology

- Understand the problem
- Gathering data
- Exploratory Data Analysis
- Data Preparation
- Classification models

Dataset

- *Public source from Kaggle*

HR Analytics: Job Change of Data Scientists

- *Size:*

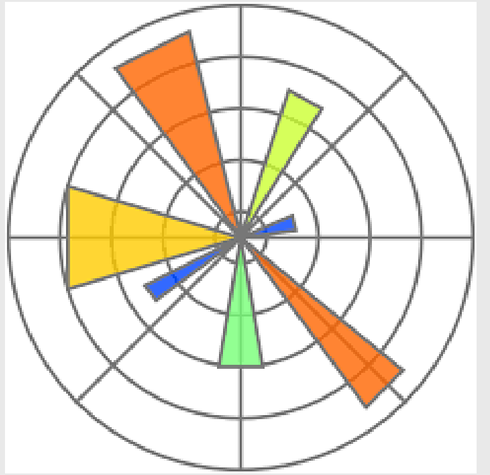
19159 records

14 columns

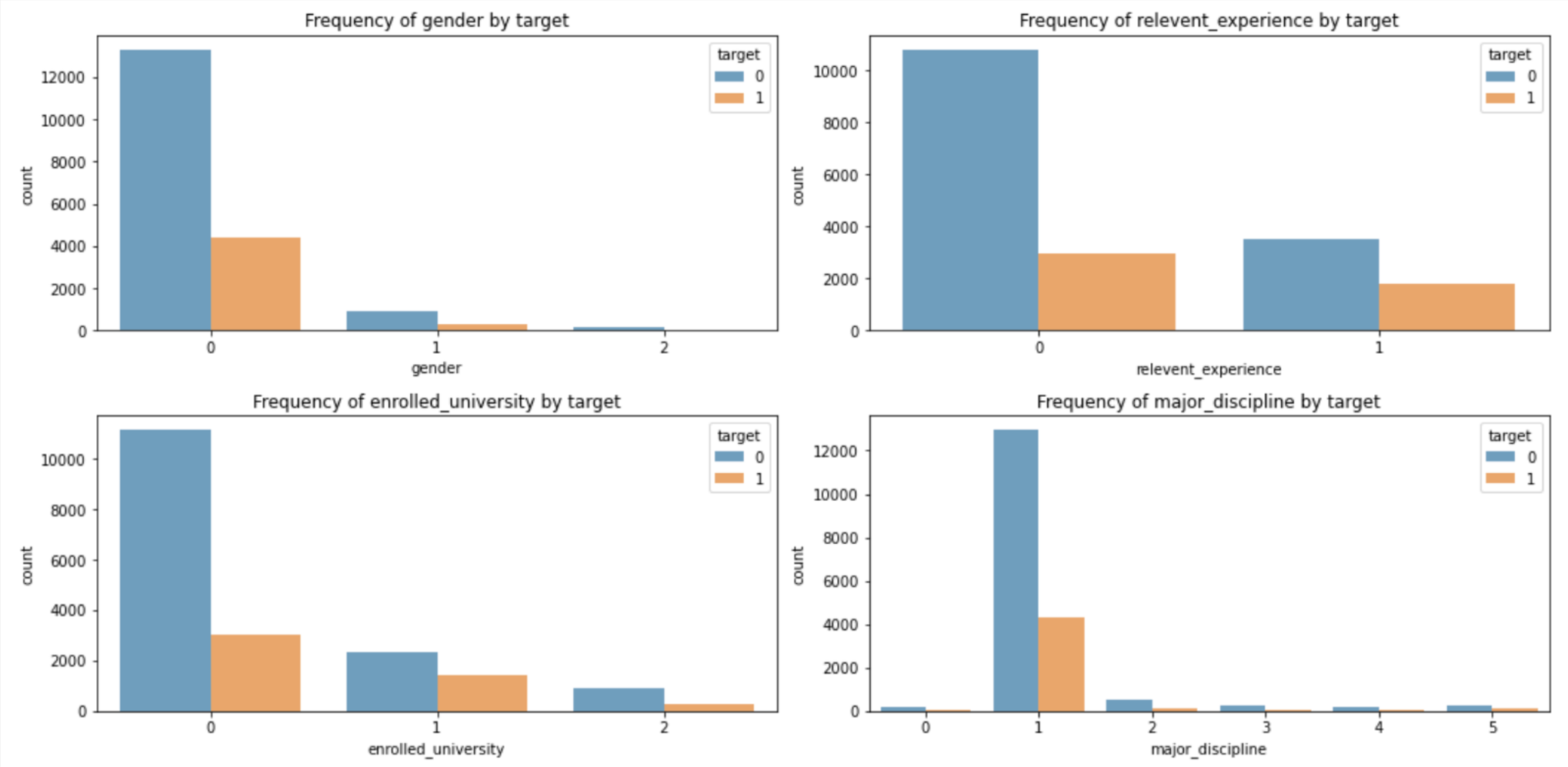
- *Target*

Leaving

Technologies and Libraries

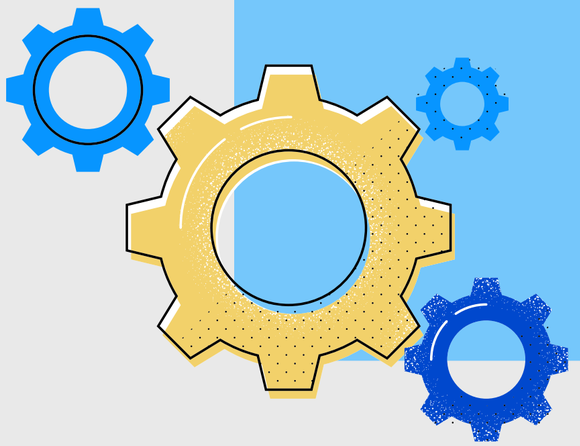


Exploratory Data Analysis



Data Preparation

- *Data Cleaning:*
  - *Replace the missing/NULL with median*
  - *Converting the datatype into numeric values*
- *Feature Selection:*
  - *Drops both the 'enrolled_id' and the 'city' columns*
- *Feature Engineering:*
  - *Encoding the columns into categorical values*
  - *Scaling using standard Scaler*
- *Imbalanced dataset:*
  - *SMOTE was use for handling the imbalanced*



Final Dataset

Dataset table:

	city_development_index	gender	relevent_experience	enrolled_university	education_level
0	0.920	0	0	0	3
1	0.776	0	1	0	3
2	0.624	0	1	1	3
3	0.789	0	1	0	3
4	0.767	0	0	0	2

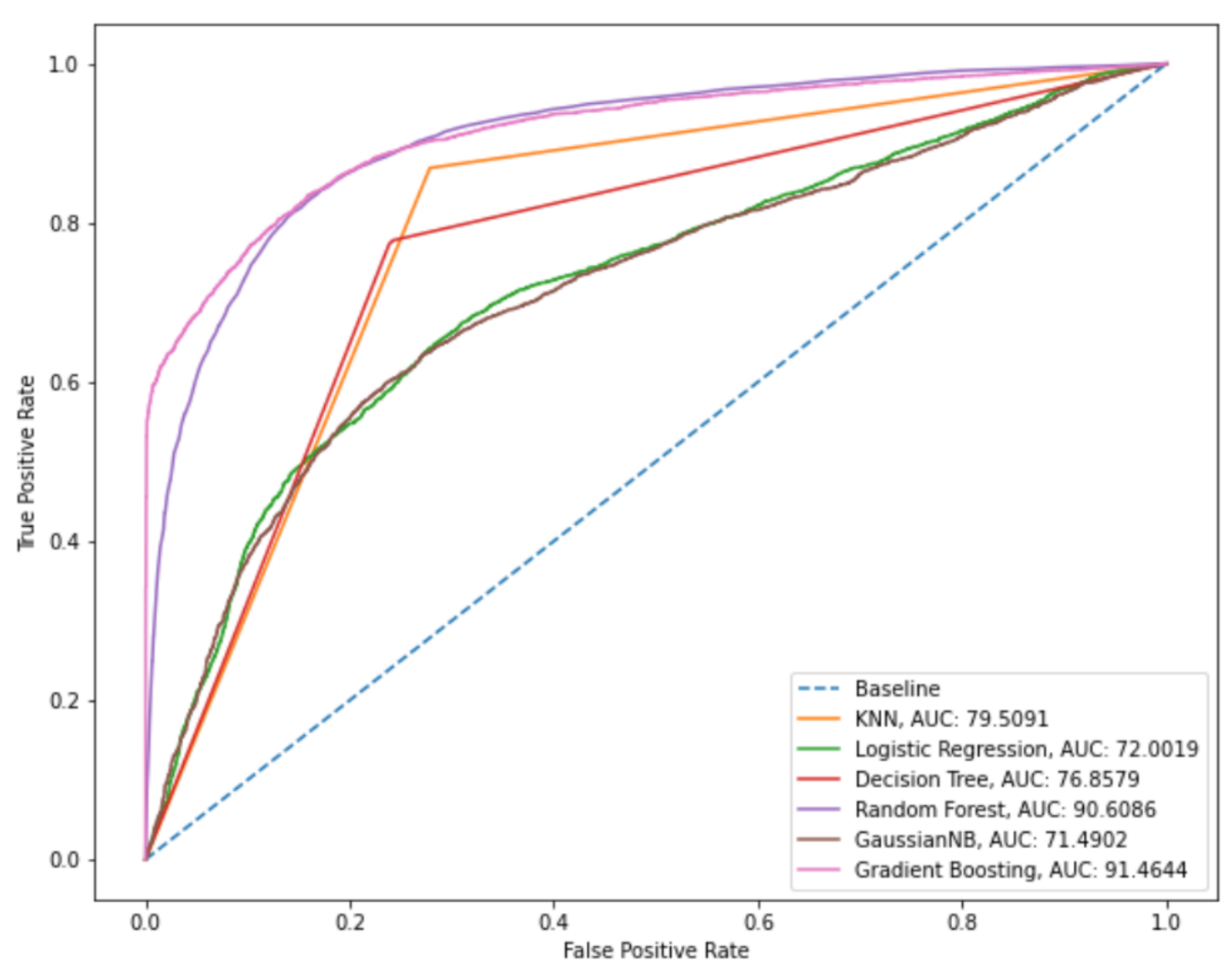
Dataset information:

```
RangeIndex: 19158 entries, 0 to 19157
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   city_development_index  19158 non-null  float64
 1   gender                  19158 non-null  Int64
 2   relevent_experience     19158 non-null  Int64
 3   enrolled_university     19158 non-null  Int64
 4   education_level         19158 non-null  Int64
 5   major_discipline        19158 non-null  Int64
 6   experience              19158 non-null  int64
 7   company_size            19158 non-null  Int64
 8   company_type            19158 non-null  Int64
 9   last_new_job            19158 non-null  Int64
 10  training_hours          19158 non-null  int64
 11  target                  19158 non-null  float64
dtypes: Int64(8), float64(2), int64(2)
```

Results and Analysis

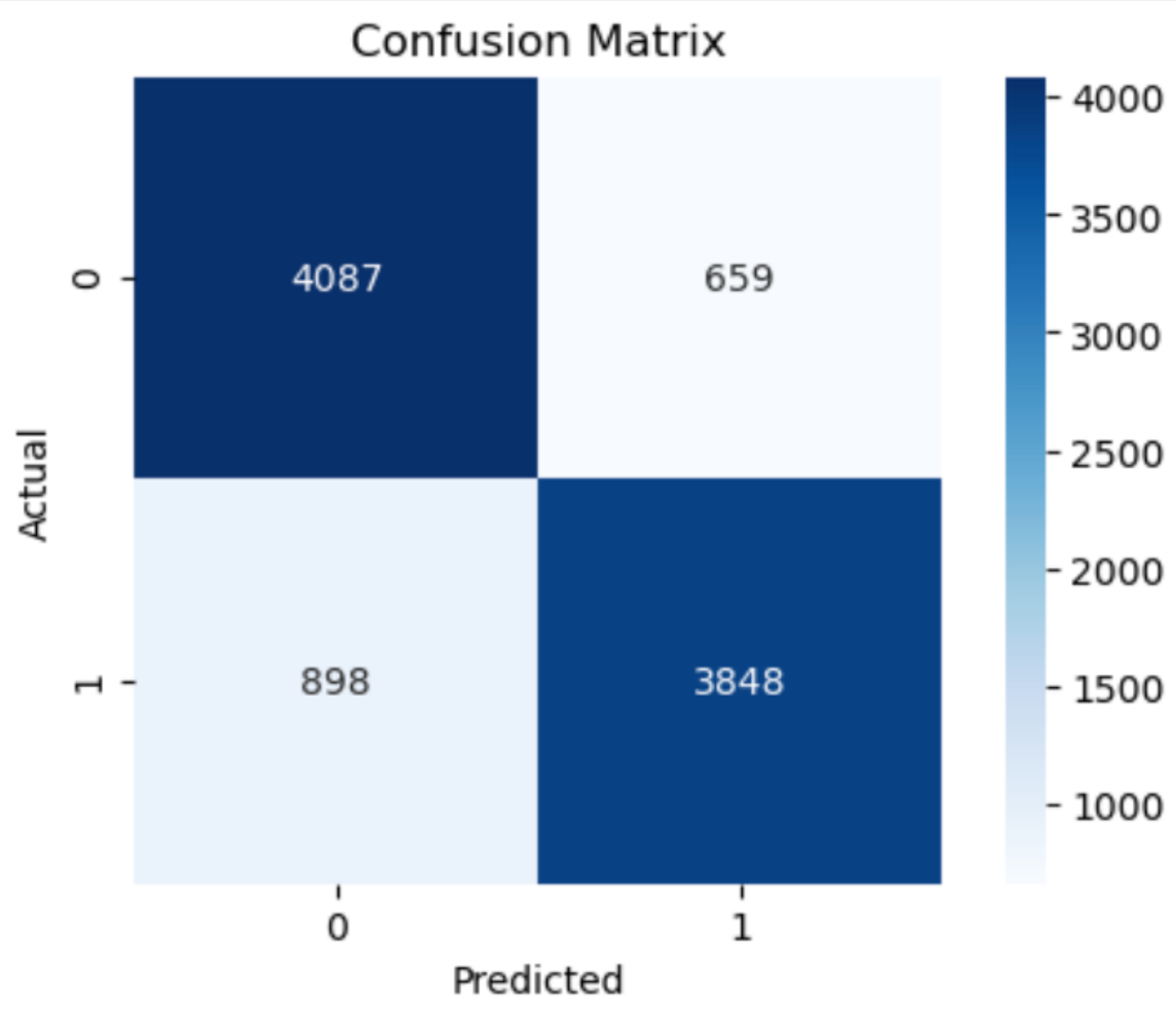
	Accuracy	F-1	Precision	Recall
KNN	0.79509	0.80914	0.75720	0.86873
Logistic Regression	0.68036	0.66348	0.70046	0.63021
Decision Tree	0.77064	0.7717	0.76799	0.77560
Random Forest	0.8350	0.83407	0.83887	0.82932
Gaussian Naive Bayes	0.66350	0.67434	0.65329	0.69679
Gradient Boosting	0.83617	0.83209	0.85337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

Results and Analysis



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

Gradient Boosting Classifier



	Gradient Boosting
Accuracy	0.81668
F-1	0.83209
Precision	0.85337
Recall	0.81184
AUC	90%

Prediction Results

	enrollee_id	Leaving
0	32403	1
1	9858	1
2	31806	1
3	27385	1
4	27724	1
5	217	1
6	21465	1
7	27302	1
8	12994	0
9	16287	1
10	10856	1
11	9272	1
12	14249	1
13	24372	1
14	14070	1
15	24914	1
16	7865	1
17	7463	0
18	21514	0
19	29033	1

Gradient Boosting was used to predict the new dataset.

Conclusion

- *Feature Engineering:*

  Encoding, Scaling

- *Imbalanced dataset:*

  SMOTE was used

- *Gradient Boosting Classifier:*

  Best ML model with AUC score 91%

Thank You!