Data Project

For HRD Bukapedia

- Promoted employee
- Top 3 best employee
- Top 3 best department

by Ranggasena T

Problem

The E-commerce company Bukapedia, which is under the auspices of Qarirlabs, already has hundreds of employees and has an average tenure of over 3 years.

Therefore, HRD wants to promote employees by providing promotion promotions and bonuses and HRD also wants to give rewards to the top 3 employees and also the departments/divisions that have the most achievements in 2022.

Data Profile

Data 2021

```
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   employee_id           40100 non-null  int64
 1   department            40100 non-null  object
 2   region                40100 non-null  object
 3   education             38332 non-null  object
 4   gender                40100 non-null  object
 5   recruitment_channel   40100 non-null  object
 6   no_of_trainings       40100 non-null  int64
 7   age                   40100 non-null  int64
 8   previous_year_rating  37076 non-null  float64
 9   length_of_service     40100 non-null  int64
 10  awards_won?           40100 non-null  int64
 11  avg_training_score    40100 non-null  int64
 12  is_promoted           40100 non-null  int64
dtypes: float64(1), int64(7), object(5)
```

Data 2021 will be the Data Train

7.5% — Data null in previous rating we fill with 3

4.4% — Data null that we dropped from education

0 — Data duplicated

0 — Outliers

Data 2022

```
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   employee_id           14707 non-null  int64
 1   department            14707 non-null  object
 2   region                14707 non-null  object
 3   education             14066 non-null  object
 4   gender                14707 non-null  object
 5   recruitment_channel   14707 non-null  object
 6   no_of_trainings       14707 non-null  int64
 7   age                   14707 non-null  int64
 8   previous_year_rating  13607 non-null  float64
 9   length_of_service     14707 non-null  int64
 10  awards_won?           14707 non-null  int64
 11  avg_training_score    14707 non-null  int64
dtypes: float64(1), int64(6), object(5)
```

We will predict from Data 2022

7.4% — Data null in previous rating we fill with 3

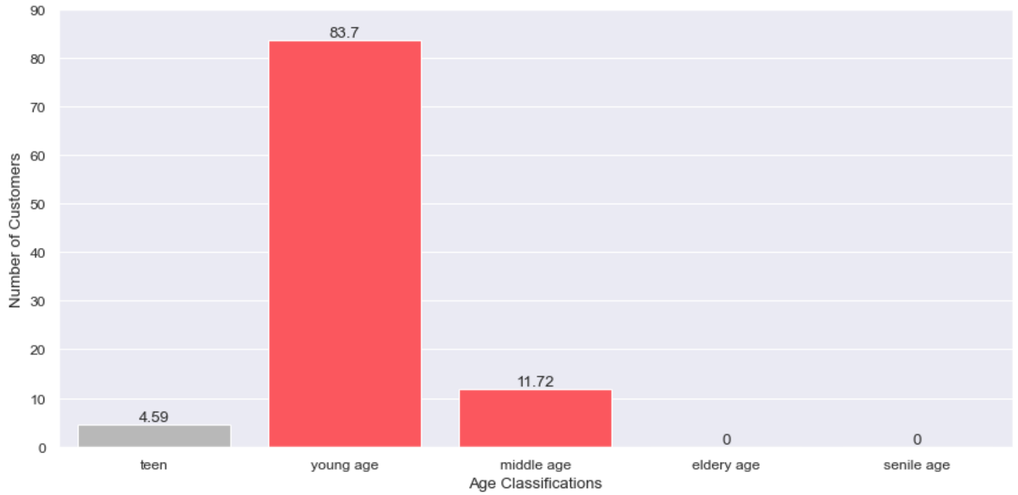
4.3% — Data null that we dropped (rows)

0 — Data duplicated

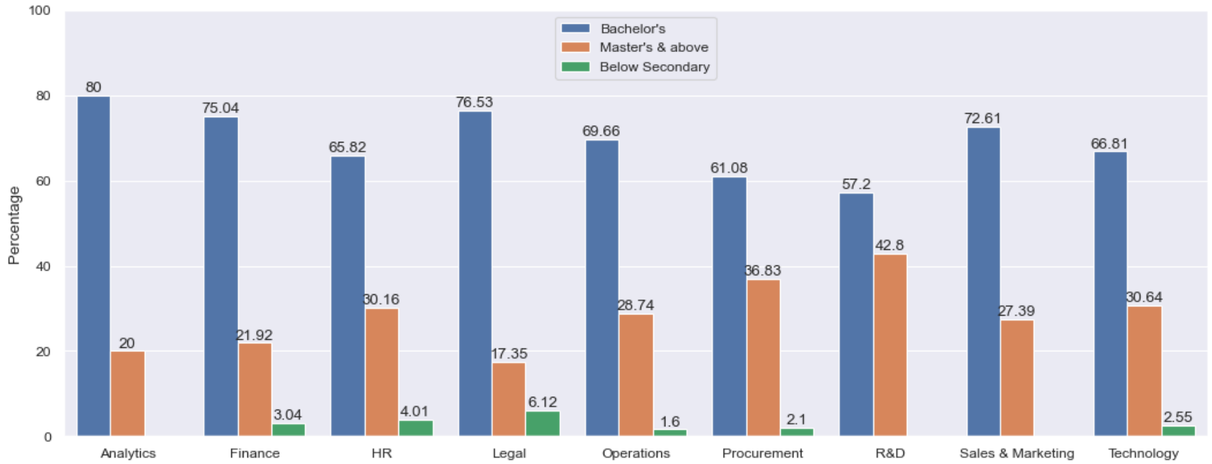
0 — Outliers

Data Insight

Data 2022



Comparison of Employee by Age (%)



Comparison of Education Background per Departement (%)

Conc =

- Generally, All departments have composition age cats are good. Young Age takes domination which are they can give a contribution of new ideas. But it has to be managed because this young age will become middle age and middle age will become elderly.
- Apart from having the most young age employees, R&D has the most Master Degree employees and none of their staff are under S1
- It needs to be explored more deeply why the Legal department has the most SMA-equivalent staff

Modeling

Data 2021

- Features Selection

Using KBest with Logistic Regression Model to choose features

```
selector.scores_
```

```
array([  25.18508376,    1.826735  ,   18.53439178,    7.64770447,
        769.75879844,    3.53786423, 1165.82729729, 1108.86819869])
```

```
X_train.columns
```

```
Index(['education', 'gender', 'no_of_trainings', 'age', 'previous_year_rating',
       'length_of_service', 'awards_won', 'avg_training_score'],
      dtype='object')
```

Result : Columns previous_year_rating, awards_won, and avg_training_score.

- Data Imbalance

We're using SMOTE to balance data and also keep the original because some algorithm already can handle data imbalance

```
y_train.value_counts()
```

```
0    28007
1     2658
```

```
y_smote.value_counts()
```

```
0    28007
1    28007
```

- Running Classification Model

Some model use two types data (original / smote) to compare

Model	SMOTE	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	Yes	0.67	0.16	0.63	0.25	0.72
Random Forest	Yes	0.52	0.12	0.68	0.2	0.62
Neural nets	Yes	0.79	0.22	0.51	0.31	0.74
K-Neighbors	Yes	0.2	0.1	0.96	0.18	0.53
XGBoost	Yes	0.16	0.1	0.98	0.17	0.64
XGBoost	No	0.92	0.73	0.12	0.21	0.73
Extra Trees	Yes	0.55	0.12	0.66	0.21	0.61
Extra Trees	No	0.92	0.69	0.12	0.21	0.73
Ligthgbm	No	0.91			0.2	

We choose XGBoost without smote (original data). The Accuracy is 92%, 73% Precission and, 0.73 AUC Score.

Predicting

Data 2022

- Predict with XGB Model

```
data_2022 = dfp[selected_cols]
```

```
cols = data_2022.columns

scaler = StandardScaler().fit(data_2022.values)
data_2022[cols] = scaler.transform(data_2022.values)
```

```
data_2022['is_promoted'] = model_xg.predict(data_2022)
```

```
data_2022.is_promoted.value_counts()
```

```
0    13835
1      231
Name: is_promoted, dtype: int64
```

```
round(231/13835*100,2)
```

```
1.67
```

1.67%

Employees in 2022 are predict to be promoted

- Finding Top 3 Employees

Using Composite Performance Index to find score. The calc is 0.61 * prev_year_rating + 0.27 * avg_training_score + 0.11 * length_of_service

	employee_id	tot_score
87	25576	499.616858
56	48831	484.961686
39	32222	477.394636

The Top Three for Employee 2022 are 25576, 48831, 32222

- Finding Top 3 Departments

department	tot_emp	emp_prom	percentage
Analytics	1434	24	1.67
Legal	296	3	1.01
Technology	1909	18	0.94

The Top Three Departments is coming from Analytics, Legal and Technology

Thank You

Contact :
ranggarain.rs@gmail.com