Employee Attrition Classification Model report

Prepared by: Lubna Almuammar

Abstract

The purpose is to support decisions that are based not on subjective aspects but on objective data analysis. The goal of this work is to analyze how objective factors influence employee attrition, in order to identify the main causes that contribute to a worker's decision to leave a company, and to be able to predict whether a particular employee will leave the company. After the training, the obtained model for the prediction of employees' attrition is tested on a real dataset provided by IBM analytics, which includes 35 features and about 1500 samples. Results are expressed in terms of classical metrics and the algorithm that produced the best results for the available dataset.

Design

- data gathering and pre-processing
- EDA & feature engineering
- model building and training

Data

- **Dataset Source** from Kaggle named "IBM HR Analytics Employee Attrition & Performance "
- **The Data Contains** records of 1470 employee with 35 features.
- **It Has Information** about employees' current employment status, the total number of companies worked for in the past. Total number of years at the current company and the current role, the education level, distance from home monthly income, etc.

Algorithm

1. **Data Collection:**
   Datasets source form kaggle.
2. **Data Understanding:**
   Understand the columns and what they represent.
3. **Data Cleaning:**
   Deleted the duplicated values. Delete null values.
4. **Data processing:**
   turn the categorical values to numeric values, dropping non-relevant features, Convert the target variable for Yes and No to 0 and 1, solve Imbalanced Data.
5. **modeling and training:**
   create a models to classify the attrition:
   a. Baseline model (KNN and Logistic Regression).

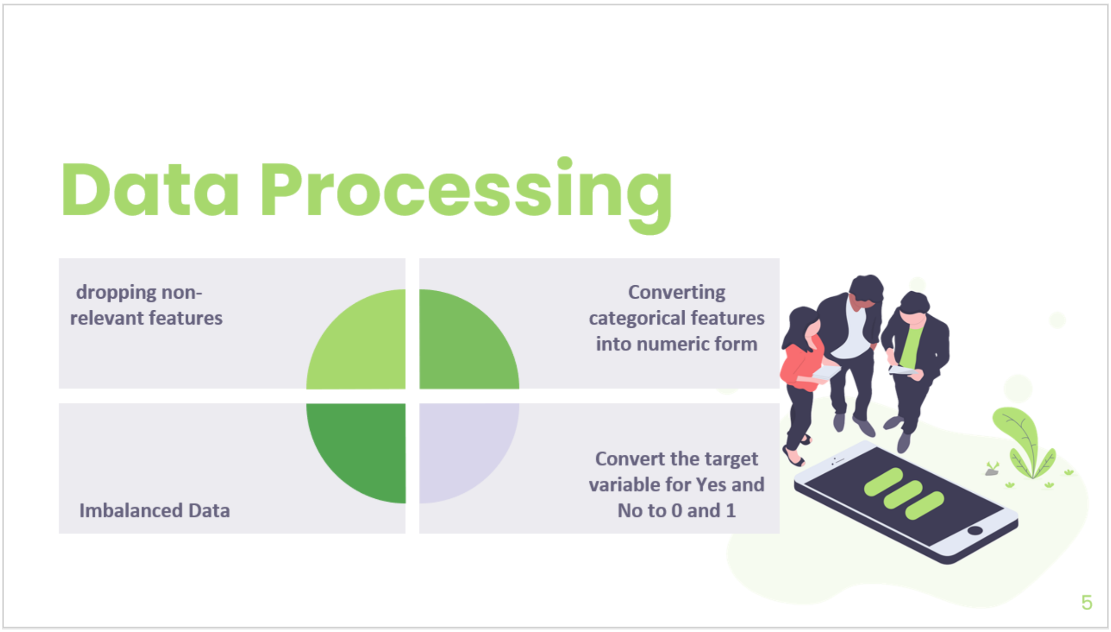
b. Logistic Regression with Down sampling and dummy variables.
c. Decision Tree Model.
d. Random Forest Classifier Model.

Tools

• **Technologies:** Jupiter notebook.

• **Libraries:** scikit-learn, Numpy, Pandas, pandas_profiling, matplotlib, seaborn, imblearn.

Communication



Data Processing

dropping non-relevant features

Converting categorical features into numeric form

Imbalanced Data

Convert the target variable for Yes and No to 0 and 1

5

Baseline model (KNN and Logistic Regression).

KNN

Training score: 0.83

Testing score:0.86

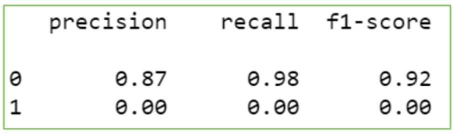
Cross_val:0.83

Logistic Regression

Training score: 0.83

Testing score:0.87

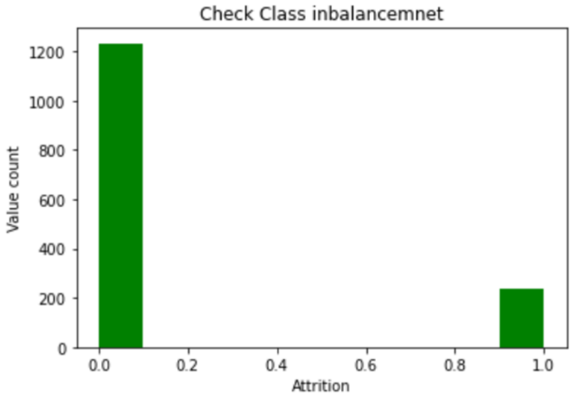
Cross_val:0.84



```
    precision     recall  f1-score

0        0.87       0.98      0.92
1        0.00       0.00      0.00
```

Class Imbalance


Check Class inbalancemnet
Value count
Attrition