BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Attrition

Group 70 - JATINDER KUMAR CHAURASIA, SWAPNIL TIWARI & RAHUL NARESHRAO SHASTRI

Data Mining

M.Tech Data Science and Engineering

13-Feb-2022

Overview

- **Objective:** Analyse and build data learning models to predict attrition.
- **Methodology:** We will build, train and test machine learning model to predict the employee attrition using attributes like distance from home, job satisfaction, years since last promotion, etc. We will be using 2 machine learning techniques – Logistic Regression and Decision Tree Classification.

Dataset

- Size of the dataset - 1470

Variable type: Variable type– Categorical, interval, binary, etc.

- Data Distribution and handling imbalanced data
- Feature Wrangling
- EDA outcomes and discussion

Feature Engineering Techniques

- Features removed: Removed Univariate Features Over18, EmployeeNumber.
- Outliers Cleaning: Replaced NAN, Nan values by the min of their feature.
- Model Optimization: Used HyperTuning technique to optimize the model performance.

Methodology

- The two machine learning techniques used
  - **Logistic Regression:**

Logistic Regression is a traditional classification algorithm involving linear discriminants. The primary output is a probability that the given input point belongs to a certain class. Based on the value of the probability, the model creates a linear boundary separating the input space into two regions. Logistic regression is easy to implement and work well on linearly separable classes, which makes it one of the most widely used classifiers.

  - **Decision Tree:**

Decision tree is a supervised method which builds classification or regression models in a tree-like structure.
The decision tree method is:

- Conceptually easy yet powerful
- Intuitive for interpretation
- Capable of handling missing values and mixed features
- Able to select variables automatically

However, its predictive power is not overly competitive. Decision tree is usually not stable with high model variance and small variations in the input data would result in a large effect on the tree structure.

Results

- Table for the evaluation metric for each ML technique used

```
LogisticRegression()
The best hyperparameters for: Logistic Regression : {'C': 0.1, 'penalty': 'l2', 'solver': 'newton-cg'} with r2 score: 0.8736632725
7
Accuracy score for Logistic Regression: 0.8785228377065112
ROC_AUC score for Logistic Regression: 0.8453332381903811
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.85      0.99      0.92       364
           1       0.76      0.21      0.33        77

    accuracy                           0.85       441
   macro avg       0.81      0.60      0.62       441
weighted avg       0.84      0.85      0.81       441

DecisionTreeClassifier()
The best hyperparameters for: Decision Tree Classifier : {'criterion': 'entropy', 'max_depth': 2} with r2 score: 0.8425716315415581
Accuracy score for Decision Tree Classifier: 0.8483965014577259
ROC_AUC score for Decision Tree Classifier: 0.7177822177822177
Classification Report for Decision Tree Classifier:
              precision    recall  f1-score   support

           0       0.87      0.94      0.90       364
           1       0.54      0.34      0.42        77

    accuracy                           0.83       441
   macro avg       0.71      0.64      0.66       441
weighted avg       0.81      0.83      0.82       441
```

Attrition

Group 70 - JATINDER KUMAR CHAURASIA, SWAPNIL TIWARI & RAHUL NARESHRAO SHASTRI

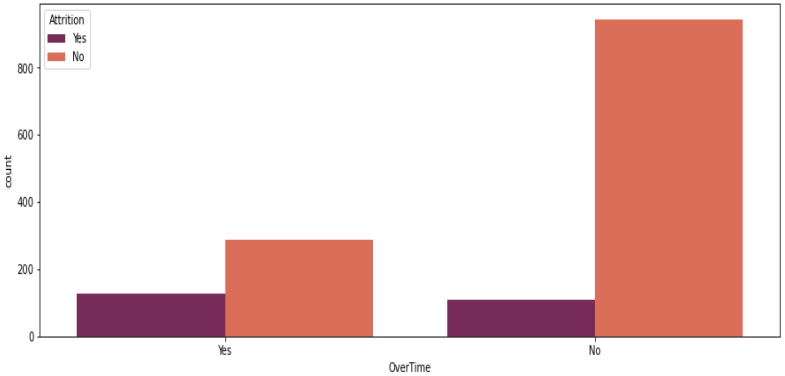
Data Mining

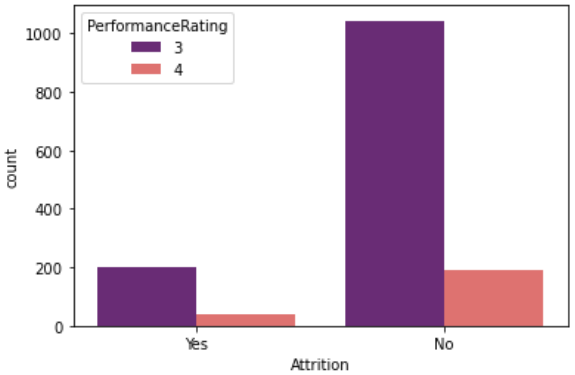
M.Tech Data Science and Engineering

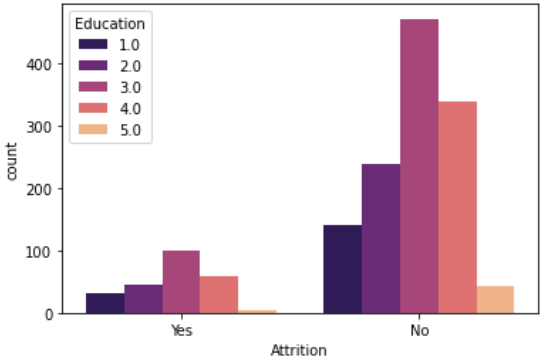
13-Feb-2022

lot of Curves:

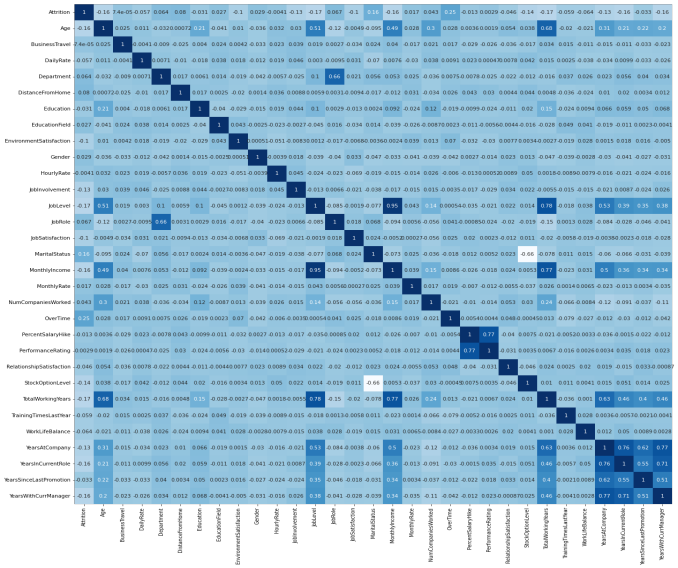
valuating Attrition against other features and how the graph changes for each feature.



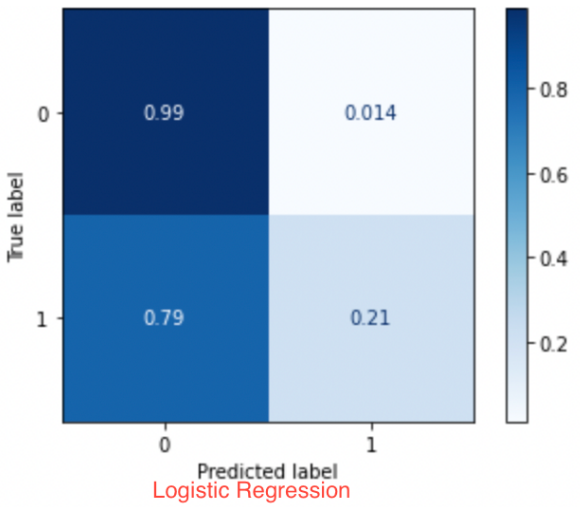


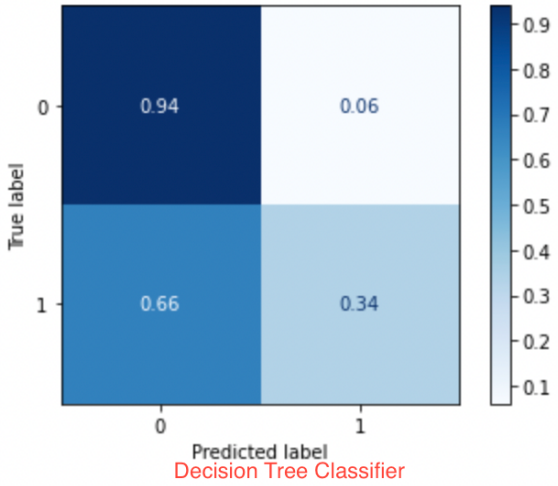


eature Correlation Heatmap Output



Logistic Regression and Decision Tree Classifier – Confusion Matrix





BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

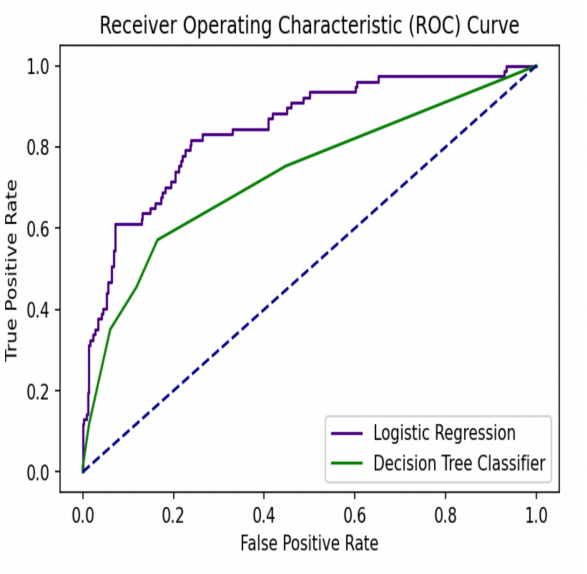
Attrition

Group 70 - JATINDER KUMAR CHAURASIA, SWAPNIL TIWARI & RAHUL NARESHRAO SHASTRI

Data Mining

M.Tech Data Science and Engineering

13-Feb-2022



	LogisticRegression	DecisionTreeClassifier
accuracy	0.878523	0.864917

Model Comparison

Conclusion: Attrition Dataset (1470*33)

1. **Data Understanding: dataset combination of Numerical and Categorical Features**
   a) **Understanding the distribution of dataset and columns.**
      a. Categorical Columns in Dataset are Education , Environment Satisfaction, JobInvolvement,JobLevel,JobSatisfaction, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, WorkLifeBalance.
      b. numerical columns are Age, DailyRate, DistanceFromHome, HourlyRate, MonthlyIncome, MonthlyRate, NumCompaniesWorked,PercentSalaryHike, TotalWorkingYears,TrainingTimesLastYear,YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion ,YearsWithCurrManager.
2. **Exploration Data Analysis:** comparing the Attrition Feature with other features and analysing the relationship b/w features , if this features is contributing to attrition.
3. **Data Cleaning and Pre-processing:**
   a) Removing Univariate Features Over18,Employee Number as they are unique for each employee and all employee are above 18.
   b) Removing Nan , NAN outliers and replacing with the column min.
   c) No null values in dataset
   d) Converted Categorical Features to their Numerical for correlation and modelling using Label Encoder
   e) Feature Correlation Using Heatmap and Pearson correlation , found that

a. Attrition has negative correlation with monthly income, total working years, Stock option Level, years at company, years in current role, years with current manager, job satisfaction, job involvement, environment satisfaction and age.
b. Attrition has positive correlation only with over time and no other parameter.
c. There is a high correlation between Percent salary hike and performance rating as well as job Level and monthly income.
d. Stock option has strong negative correlation with Marital Status.
e. Years at company, years in current role, years since last promotion, years with current manager, total working years, monthly income and job level are also correlated.

4. Modelling:

a) Using Logistic Regression and Decision Tree Classifier models to predict the attrition.
b) Use Variance Threshold to select the features
c) To improve the Model performance used Hyper Tuning(a technique used for choosing a set of optimal hyperparameters for a learning algorithm) used StratifiedKFold(making k folds of the Dataset) and GridSearchCV (to find best parameters for Learning Model Technique) during model learning.
d) After comparing those two models – found that Logistic Regression do have better accuracy than Decision Tree Classifier.
e) Logistic Regression seems to be better model for predicting the Attrition.