**DATA SCIENCE PROJECT**

**ON**

**ATTRITION DATA**

**SNEHAL JAGTAP**

**ETL HIVE**

**DATA SUMMARY**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Column Name** | **Description** |
| 1 | Age | Age of the employee |
| 2 | Attrition | Whether left organization (1=Left, 0=Not left) |
| 3 | BusinessTravel | Rarely, frequently, no travel |
| 4 | DailyRate | USD per day |
| 5 | Department | R&D, HR, Sales |
| 6 | DistanceFromHome | KM |
| 7 | Education | 1=Lowest (12th Std), 5 = Highest (PhD) |
| 8 | EducationField | Life science, HR, Medical, Marketing, Technical degree, Others |
| 9 | EmployeeCount | (Count =1 ) May be ignored |
| 10 | EmployeeNumber | Emp Id |
| 11 | EnvironmentSatisfaction | (1=Lowest, 4=Highest) |
| 12 | Gender | Male, Female |
| 13 | HourlyRate |  |
| 14 | JobInvolvement | (1=Lowest, 4=Highest) |
| 15 | JobLevel | Level in Heirarchy(1=Lowest, 4=Highest) |
| 16 | JobRole | Job Designation |
| 17 | JobSatisfaction | (1=Lowest, 4=Highest) |
| 18 | MaritalStatus | Single, Married, Divorced |
| 19 | MonthlyIncome | income |
| 20 | MonthlyRate | rate |
| 21 | NumCompaniesWorked | Number of Previous Companies worked |
| 22 | Over18 | Adult (Yes,No) |
| 23 | OverTime | Does Overtime (Yes, No) |
| 24 | PercentSalaryHike | % |
| 25 | PerformanceRating | Last Performance Rating (1=Lowest,5=Highest) |
| 26 | RelationshipSatisfaction | Relationship Satisfaction within company (1=Lowest, 4=Highest) |
| 27 | StandardHours | Standatrd Working hours in a fortnight (2 weeks) |
| 28 | StockOptionLevel | 0=No option, 1 = Low, 2 = Medim, 3 = High |
| 29 | TotalWorkingYears | Total Experience |
| 30 | TrainingTimesLastYear | Number of days of Training attended in last yesr |
| 31 | WorkLifeBalance | 1=Lowest, 4 = Highest |
| 32 | YearsAtCompany | Number of years with current company |
| 33 | YearsInCurrentRole | Number of years in current role |
| 34 | YearsSinceLastPromotion | Number of years since last promotion |
| 35 | YearsWithCurrManager | Number of years with current manager |

* Assuming target as attrition, form Classification Models and choose the best model.

**STEPS OF THE PROJECT**

* Explore the Data using Exploratory Data Analysis –For all predictors and target
  + - For univariate analysis
      * + Quantitative data :

Box Plot

Histogram

Histogram with distribution of data

Violin Plot

Strip plot

* + - * + Categorical data:

Pie chart

Count Plot

Bar Chart

* + - For bivariate analysis
      * + Categorical 🡪 Categorical

Contengency table

* + - * + Categorical 🡪 Quantitative

Side by side box plot

* + - * + Quantitative 🡪 Quantitative

Scatter plot

* + - * + Quantitative 🡪 Categorical

Horizontal side by side box plot

* **Test of hypothesis**

We get a p-value by TOH , if it is in between 0.05 – 0.1 we include that predictor in model as target is dependent on that predictor.

* + - * + Categorical 🡪 Categorical

Chi square test

* + - * + Categorical 🡪 Quantitative

Two sample t test (2 categories – independent sample)

Anova F test (>2 categories – independent samples)

Paired t test (2 categories – dependent samples)

* + - * + Quantitative 🡪 Quantitative

T test for slope (using least square estimate)

* + - * + Quantitative 🡪 Categorical

Logistic regression

* Select the predictors by TOH.
* For quantitative predictors check outliers, if outliers are present, remove the outliers.
* For categorical predictors, convert them into dummy variables for further processing.
* By performing all the above steps , respective files have been created

Attrition\_EDA\_TOH

MonthlyIncome\_EDA\_TOH

Above mentioned files are used for classification and regression models respectively.

**Predictors / Features selected by TOH**

|  |
| --- |
| **Attrition** |
| Attrition |
| Age |
| BusinessTravel |
| DailyRate |
| Department |
| DistanceFromHome |
| EducationField |
| EnvironmentSatisfaction |
| JobInvolvement |
| JobLevel |
| JobRole |
| JobSatisfaction |
| MaritalStatus |
| MonthlyIncome |
| NumCompaniesWorked |
| OverTime |
| RelationshipSatisfaction |
| StockOptionLevel |
| TotalWorkingYears |
| TrainingTimesLastYear |
| WorkLifeBalance |
| YearsAtCompany |
| YearsInCurrentRole |

**CLASSIFICATION MODELS**

1. **Logistic Regression**:

Train accuracy = 0.8901734104046243

Test accuracy = 0.8467741935483871

K-Fold cross validation = 0.878588612670409

1. **K- nearest Neighbor (KNN)**:

Train accuracy = 0.8508670520231214

Test accuracy = 0.8118279569892473

K-Fold cross validation = 0.8346902937420179

1. **Naïve Base:**

Gaussian Naïve base

Accuracy : 0.3010752688172043

Precision : 0.21406727828746178

Recall : 0.958904109589041

f1\_score : 0.35

Bernoulli’s Naïve base

Accuracy : 0.8413978494623656

Precision : 0.6166666666666667

Recall : 0.5068493150684932

f1\_score : 0.5563909774436091

1. **Classification trees:**

Decision Tree Classifier on training set: 0.8543352601156069

Decision Tree Classifier on test set: 0.8118279569892473

Bagging classifier on training set: 0.8635838150289017

Bagging classifier on test set: 0.8118279569892473

OOB classifier on test set: 0.8497109826589595

Random Forest classifier on training set: 0.8775

Random Forest classifier on test set: 0.8145

Random Forest OOB classifier on test set: 0.8405

Gradient Boosting classifier on training set: 0.8578

Gradient Boosting classifier on test set: 0.8145

**BEST MODELS**

**Prediction of Attrition [classification models]**

**Logistic Regression** model have best accuracy

Training : approx. 89

Testing : approx. 85

K- Fold Cross validation : approx. 88

**PCA model** :

Training : approx. 89

Testing : approx. 84

K- Fold Cross validation : approx. 87