**DATA SCIENCE PROJECT**

**ON**

**MONTHLY INCOME**

**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 Monthly Income, form Regression 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 file has been created

MonthlyIncome\_EDA\_TOH

Above mentioned files are used for regression models respectively.

**Predictors / Features selected by TOH**

|  |
| --- |
| **Monthly Income** |
| Age |
| Department |
| Education |
| EducationField |
| JobLevel |
| JobRole |
| MaritalStatus |
| MonthlyIncome |
| NumCompaniesWorked |
| StockOptionLevel |
| TotalWorkingYears |
| YearsAtCompany |
| YearsInCurrentRole |
| YearsSinceLastPromotion |
| YearsWithCurrManager |

**REGRESSION MODELS**

1. **Multiple linear regression :**

Train accuracy

r2 = 0.8265282456562424

MSE = 1180506.6133452668

Test accuracy

r2 = 0.8161659207944133

MSE = 1271589.7444297585

Kfold cross validation

r2 = 0.809411085507314

1. **Ridge regression:**

Train accuracy = 0.8260625860505242

Test accuracy = 0.8169636819305401

1. **Lasso regression**:

Train accuracy = 0.8264100777266908

Test accuracy = 0.8158681968901649

1. **Regression Trees:**

Decision Tree Classifier on training set: 0.057623049219687875

Decision Tree Classifier on test set: 0.002793296089385475

Bagging classifier on training set: 0.7707082833133253

Bagging classifier on test set: 0.002793296089385475

OOB classifier on test set: 0.0

Random Forest classifier on training set: 0.8259

Random Forest classifier on test set: 0.0028

Random Forest OOB classifier on test set: 0.0024

**BEST MODELS**

**Prediction of Monthly Income [Regression models]**

All the models have approximately same accuracies (MLR, Ridge, Lasso)

Training : approx. 82.60

Testing : approx. 81.50