**Dataset for Problem 2:**[**Holiday\_Package.csv**](https://olympus.mygreatlearning.com/courses/73747/files/4760225/download?wrap=1)

**Data Dictionary:**

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Holiday\_Package | Opted for Holiday Package yes/no? |
| Salary | Employee salary |
| age | Age in years |
| edu | Years of formal education |
| no\_young\_children | The number of young children (younger than 7 years) |
| no\_older\_children | Number of older children |
| foreign | foreigner Yes/No |

1. The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, especially identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have a symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. For this is a classification problem, the dependence of the response on the predictors needs to be investigated.

Solution-

* Lets check the shape of the data-

(872, 8)

We can see that our data set is having 26967 rows and 10 columns.

* Lets check the information –

RangeIndex: 872 entries, 0 to 871

Data columns (total 8 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 872 non-null int64

1 Holliday\_Package 872 non-null object

2 Salary 872 non-null int64

3 age 872 non-null int64

4 educ 872 non-null int64

5 no\_young\_children 872 non-null int64

6 no\_older\_children 872 non-null int64

7 foreign 872 non-null object

dtypes: int64(6), object(2)

memory usage: 54.6+ KB

We can observe that there is 6 intergers and 2 object type data type in columns.

* Let check the data describtion-

| **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Unnamed: 0** | 872.0 | 436.500000 | 251.869014 | 1.0 | 218.75 | 436.5 | 654.25 | 872.0 |
| **Salary** | 872.0 | 47729.172018 | 23418.668531 | 1322.0 | 35324.00 | 41903.5 | 53469.50 | 236961.0 |
| **age** | 872.0 | 39.955275 | 10.551675 | 20.0 | 32.00 | 39.0 | 48.00 | 62.0 |
| **educ** | 872.0 | 9.307339 | 3.036259 | 1.0 | 8.00 | 9.0 | 12.00 | 21.0 |
| **no\_young\_children** | 872.0 | 0.311927 | 0.612870 | 0.0 | 0.00 | 0.0 | 0.00 | 3.0 |
| **no\_older\_children** | 872.0 | 0.982798 | 1.086786 | 0.0 | 0.00 | 1.0 | 2.00 | 6.0 |

* Lets analyze more about object type columns –

Number of yes and no in Holliday\_Package-

no 471

yes 401

Number of yes and no in foreign-

no 656

yes 216

Number of unique value counts in the column name no\_older\_children-

0 393

2 208

1 198

3 55

4 14

5 2

6 2

Number of unique value counts in the column name no\_older\_children-

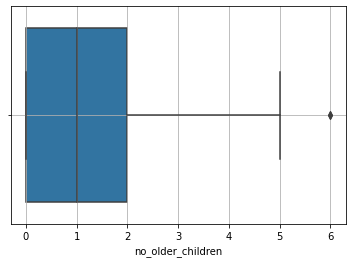
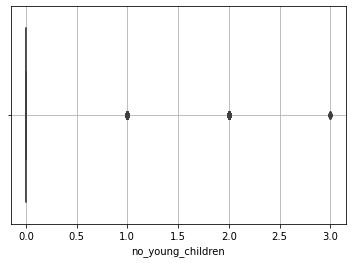
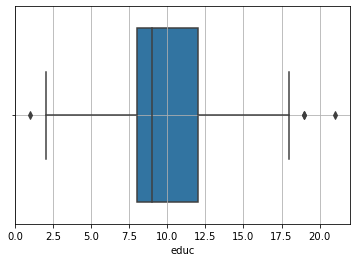
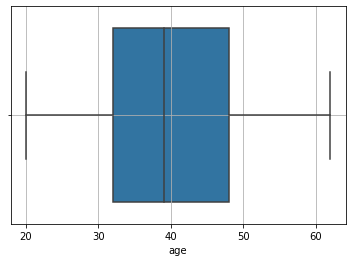
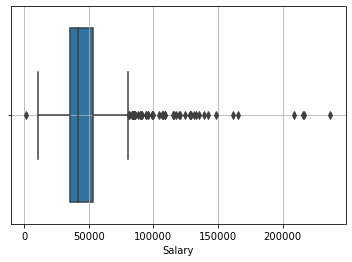
0 665

1 147

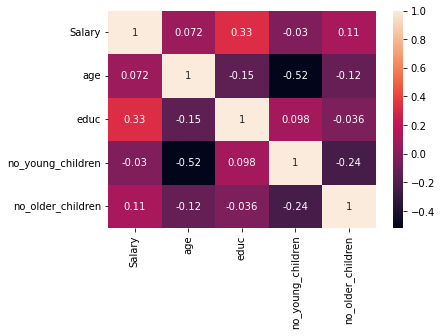
2 55

3 5

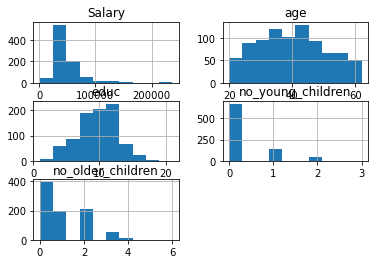
**Let’s analyse about outliers by using box plot-**



* **Lets check the correlation in the data type-**



**Let see the distribution-**



**2- The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, especially identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have a symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. For this is a classification problem, the dependence of the response on the predictors needs to be investigated.**

**Solution-**

Model 1- With all the 6 dependent variables

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | Holliday\_Package\_yes | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 865 |
| **Method:** | MLE | **Df Model:** | 6 |
| **Date:** | Fri, 30 Sep 2022 | **Pseudo R-squ.:** | 0.1281 |
| **Time:** | 07:24:31 | **Log-Likelihood:** | -524.53 |
| **converged:** | True | **LL-Null:** | -601.61 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 1.023e-30 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 2.3259 | 0.554 | 4.199 | 0.000 | 1.240 | 3.411 |
| **Salary** | -1.814e-05 | 4.35e-06 | -4.169 | 0.000 | -2.67e-05 | -9.61e-06 |
| **age** | -0.0482 | 0.009 | -5.314 | 0.000 | -0.066 | -0.030 |
| **educ** | 0.0392 | 0.029 | 1.337 | 0.181 | -0.018 | 0.097 |
| **no\_young\_children** | -1.3173 | 0.180 | -7.326 | 0.000 | -1.670 | -0.965 |
| **no\_older\_children** | -0.0204 | 0.074 | -0.276 | 0.782 | -0.165 | 0.124 |
| **foreign\_yes** | 1.3216 | 0.200 | 6.601 | 0.000 | 0.929 | 1.714 |

|  |  |
| --- | --- |
| **Pseudo R-squ.:** | 0.1281 |
| **R-squ.:** | 0.1100 |

Lets check multicolinarity in the model-

Salary VIF = 1.17

age VIF = 1.58

educ VIF = 1.4

no\_young\_children VIF = 1.57

no\_older\_children VIF = 1.19

foreign\_yes VIF = 1.27

As we can see from the box plot there are so many outliers in salary data set, So will transform the outliers-

**Model 2-**

Lets build the model 2 with following variables-

M\_2 = 'Holliday\_Package\_yes~Salary+age+educ+no\_young\_children+no\_older\_children+foreign\_yes'

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | Holliday\_Package\_yes | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 865 |
| **Method:** | MLE | **Df Model:** | 6 |
| **Date:** | Fri, 30 Sep 2022 | **Pseudo R-squ.:** | 0.1244 |
| **Time:** | 07:28:37 | **Log-Likelihood:** | -526.78 |
| **converged:** | True | **LL-Null:** | -601.61 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 9.138e-30 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 2.5432 | 0.559 | 4.550 | 0.000 | 1.448 | 3.639 |
| **Salary** | -2.088e-05 | 5.26e-06 | -3.970 | 0.000 | -3.12e-05 | -1.06e-05 |
| **age** | -0.0496 | 0.009 | -5.491 | 0.000 | -0.067 | -0.032 |
| **educ** | 0.0342 | 0.029 | 1.172 | 0.241 | -0.023 | 0.091 |
| **no\_young\_children** | -1.3287 | 0.180 | -7.386 | 0.000 | -1.681 | -0.976 |
| **no\_older\_children** | -0.0251 | 0.074 | -0.341 | 0.733 | -0.169 | 0.119 |
| **foreign\_yes** | 1.3037 | 0.200 | 6.519 | 0.000 | 0.912 | 1.696 |

|  |  |
| --- | --- |
| **Pseudo R-squ.:** | 0.1244 |
| **Adj R-squ.:** | 0.1144 |

**Model 3-**

M\_3 = 'Holliday\_Package\_yes~Salary+age+educ+no\_young\_children+foreign\_yes'

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | Holliday\_Package\_yes | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 866 |
| **Method:** | MLE | **Df Model:** | 5 |
| **Date:** | Fri, 30 Sep 2022 | **Pseudo R-squ.:** | 0.1243 |
| **Time:** | 07:29:07 | **Log-Likelihood:** | -526.84 |
| **converged:** | True | **LL-Null:** | -601.61 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 1.671e-30 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 2.4783 | 0.525 | 4.724 | 0.000 | 1.450 | 3.506 |
| **Salary** | -2.117e-05 | 5.19e-06 | -4.079 | 0.000 | -3.13e-05 | -1.1e-05 |
| **age** | -0.0487 | 0.009 | -5.677 | 0.000 | -0.065 | -0.032 |
| **educ** | 0.0351 | 0.029 | 1.209 | 0.227 | -0.022 | 0.092 |
| **no\_young\_children** | -1.3080 | 0.169 | -7.747 | 0.000 | -1.639 | -0.977 |
| **foreign\_yes** | 1.3028 | 0.200 | 6.517 | 0.000 | 0.911 | 1.695 |

|  |  |
| --- | --- |
| **Pseudo R-squ.:** | 0.1243 |
| **Adjust R-squ.:** | 0.1159 |

**Model 4-**

M\_4 = 'Holliday\_Package\_yes~Salary+age+no\_young\_children+foreign\_yes'

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| **Dep. Variable:** | Holliday\_Package\_yes | **No. Observations:** | 872 |
| **Model:** | Logit | **Df Residuals:** | 867 |
| **Method:** | MLE | **Df Model:** | 4 |
| **Date:** | Fri, 30 Sep 2022 | **Pseudo R-squ.:** | 0.1231 |
| **Time:** | 07:29:29 | **Log-Likelihood:** | -527.58 |
| **converged:** | True | **LL-Null:** | -601.61 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 5.267e-31 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | 2.8128 | 0.448 | 6.282 | 0.000 | 1.935 | 3.690 |
| **Salary** | -1.932e-05 | 4.94e-06 | -3.911 | 0.000 | -2.9e-05 | -9.64e-06 |
| **age** | -0.0504 | 0.008 | -5.962 | 0.000 | -0.067 | -0.034 |
| **no\_young\_children** | -1.3023 | 0.169 | -7.707 | 0.000 | -1.633 | -0.971 |
| **foreign\_yes** | 1.2092 | 0.183 | 6.592 | 0.000 | 0.850 | 1.569 |

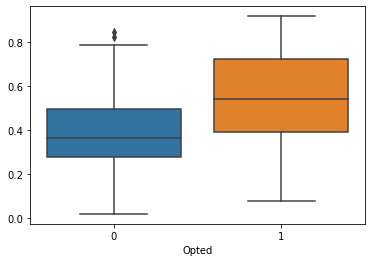
|  |  |
| --- | --- |
| **Pseudo R-squ.:** | 0.1231 |
| **Adjust R-squ.:** | 0.1164 |

**Conclusion of problem 2-After comparing all the 4 models based on their adjusted R square score model 4 comes out as best model.**

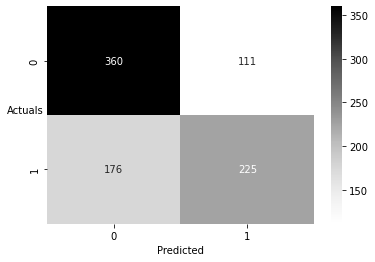
**3)- Alternatively, if prediction accuracy of the full scholarship is the only objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (2) to compare accuracy in training and test sets. Compare the final model of Part (2) and the proposed one in Part (3). Which model provides the most accurate prediction? If the model found in Part (2) is different from the proposed model in Part (3), give an explanation.**

**Solution-**

* Graphical distribution of target variable-



Confusion matrix for actual target variable and predicted target variable-



True Negative: 360

False Positives: 111

False Negatives: 176

True Positives: 225

**Let us now go ahead and print the classification report to check the various other parameters.**

precision recall f1-score support

0 0.672 0.764 0.715 471

1 0.670 0.561 0.611 401

accuracy 0.671 872

macro avg 0.671 0.663 0.663 872

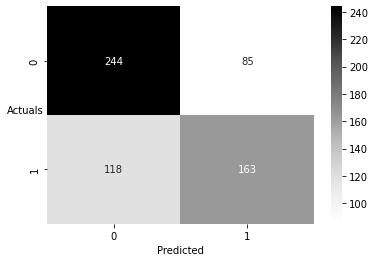
weighted avg 0.671 0.671 0.667 872

Now lest devide the data set into train & test with all the 6 independent variables given below-

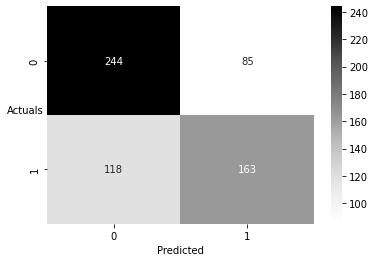
'Salary', 'age', 'educ', 'no\_young\_children',

'no\_older\_children', 'foreign\_yes'

Now lets check confusion matrix for training data for model one-

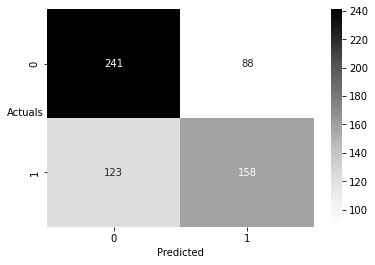


Now lets check confusion matrix for test data for model one-

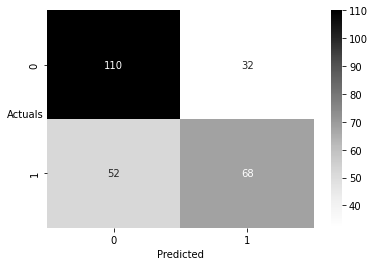


## using model4 to build the model on training data and on predect for traing and test data

Now lets check confusion matrix for training data for model four-



Now lets check confusion matrix for test data for model one-



Conclusion for problem 3-

Accuracy Score - Model 1

Training: 0.6672131147540984

Testing: 0.648854961832061

Accuracy Score - Model 4

Training: 0.6540983606557377

Testing: 0.6793893129770993

**As we can observe that Model 4 is performing better than Model 1, although model 4 is also not very accurate but in comparison to other models its better than others.**