PETER GACHUKI CASE STUDY REPORT

METHODOLOGY

The dataset was thoroughly explored with the help of pandas and numpy libraries. The missing values in the dataset were detected and handled: for some, the rows were dropped and for others, they were replaced by filling them with appropriate values. Categorical data from the dataset was handled by encoding them so as to enable the machine to process it, as compared to string categories. The dataset was then split into training and testing sets using scikit-learn library.Various classification models(Logistic regression, decision trees, gradient boosting classifier, support vector classifier, K-Nearest-Neighbours and Naive Bayes Classifier) were used and their performance metrics analysed. The evaluation metrics used for the models include precision, recall, F1-score and confusion matrices.

KEY FINDINGS

- The handling of categorical features and missing values significantly impacted the models' performance
- The gender is imbalanced especially in the gender column with an overwhelming representation of Male as compared to the other
- The support vector classifier and gradient boost classifier performed best both with an overall accuracy of 85%

MODEL PERFORMANCE METRICS

Logistic Regression Model:
Precision: High for 0 and moderate for 1
Recall: High for 0 and low for 1
F1-score: High for 0 and low for 1
Accuracy: 84%

Decision Trees Model:
Precision: High for 0 and low for 1
Recall: High for 0 and low for 1
F1-score: High for 0 and low for 1
Accuracy: 77%

Gradient Boosting Model:
Precision: High for 0 and moderate for 1
Recall: High for 0 and low for 1
F1-score: High for 0 and low for 1
Accuracy: 85%

Support Vector Model:
Precision: High for 0 and moderate for 1
Recall: High for 0 and moderate for 1

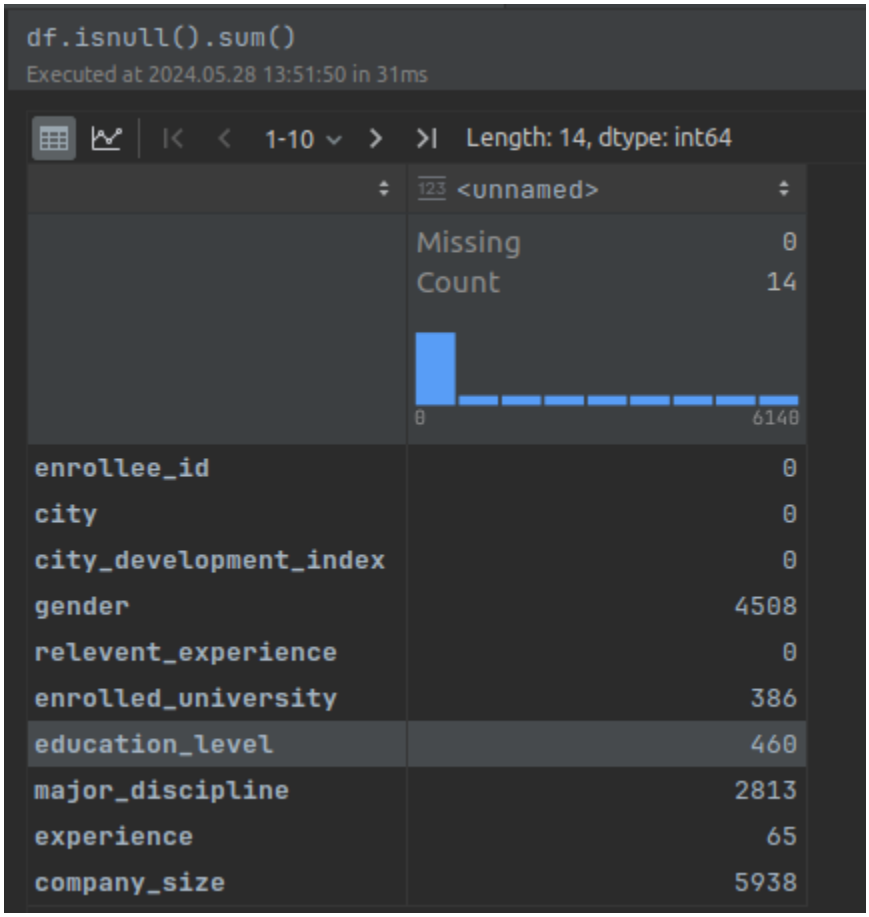
F1-score: High for 0 and moderate for 1
Accuracy: 85%

K-Nearest Neighbours Model:
Precision: High for 0 and moderate for 1
Recall: High for 0 and low for 1
F1-score: High for 0 and moderate for 1
Accuracy: 82%

Naive Bayes Model:
Precision: High for 0 and moderate for 1
Recall: High for 0 and moderate for 1
F1-score: High for 0 and moderate for 1
Accuracy: 82%

The 2 best models(Support Vector Classifier and Gradient Boosting Classifier were saved using the pickle library)

Preprocessing



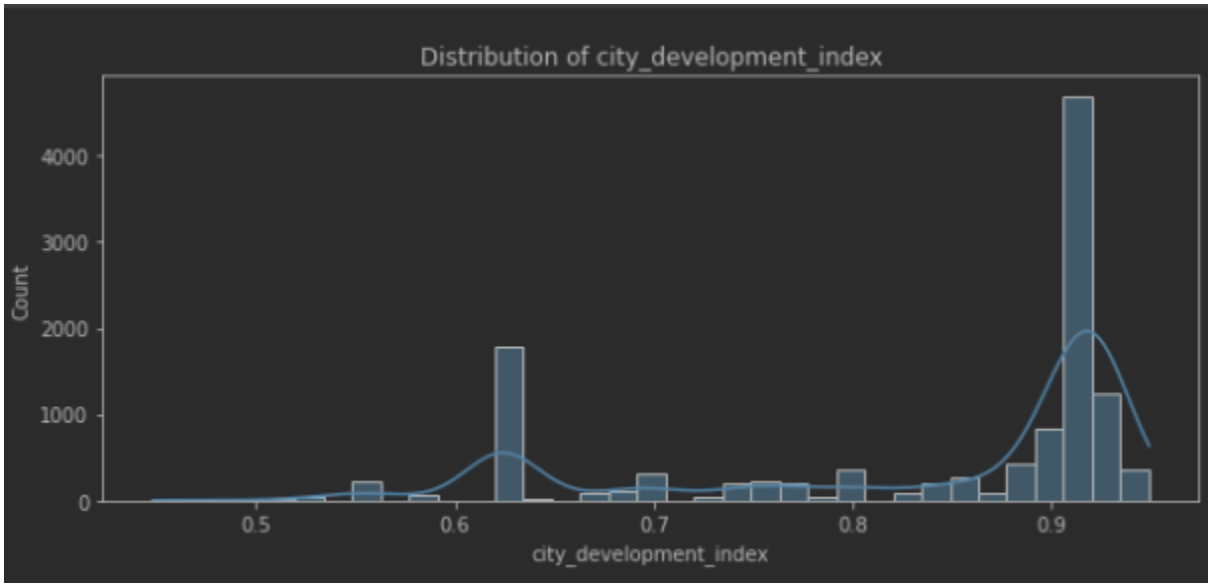
The above shows the number of null values per respective column. For all but the gender columns, the rows containing the null values were dropped. For the gender column, the null values were filled in proportionally in accordance to their share in the dataset.

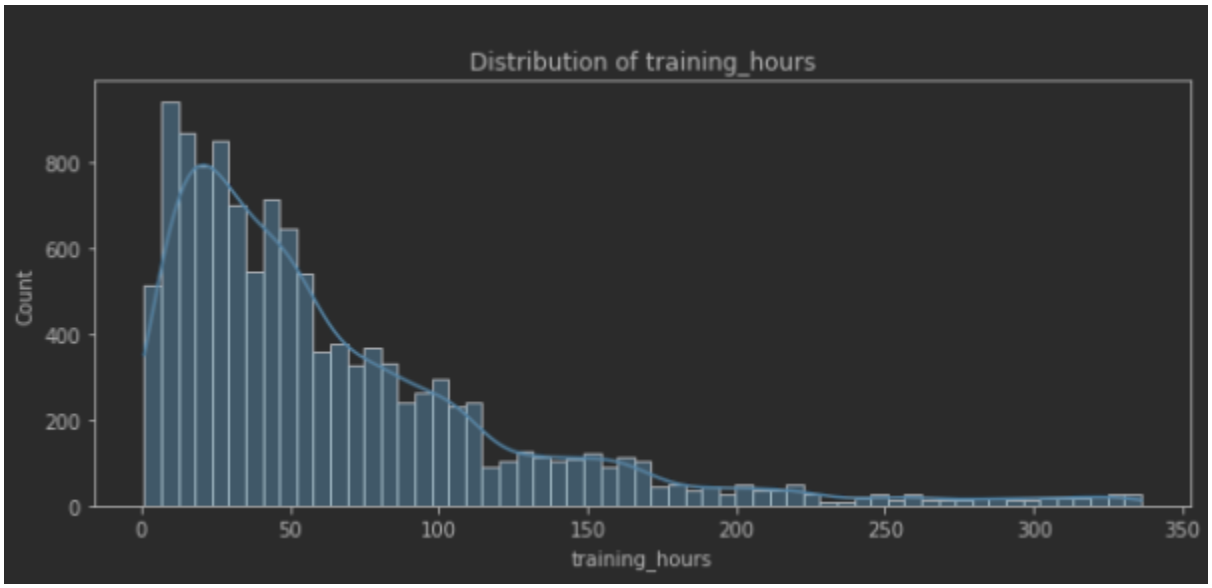
```
df["experience"].unique()
Executed at 2024.05.28 13:51:53 in 9ms

array(['15', '>20', '13', '7', '17', '5', '16', '1', '2', '11', '<1',
       '14', '18', '19', '12', '10', '6', '9', '3', '4', '8', '20'],
      dtype=object)
```

For the experience column, >20 and <1 were replaced before feeding the data to the model.Similarly, company_sizea and last_new_job had ranges which were duly replaced.

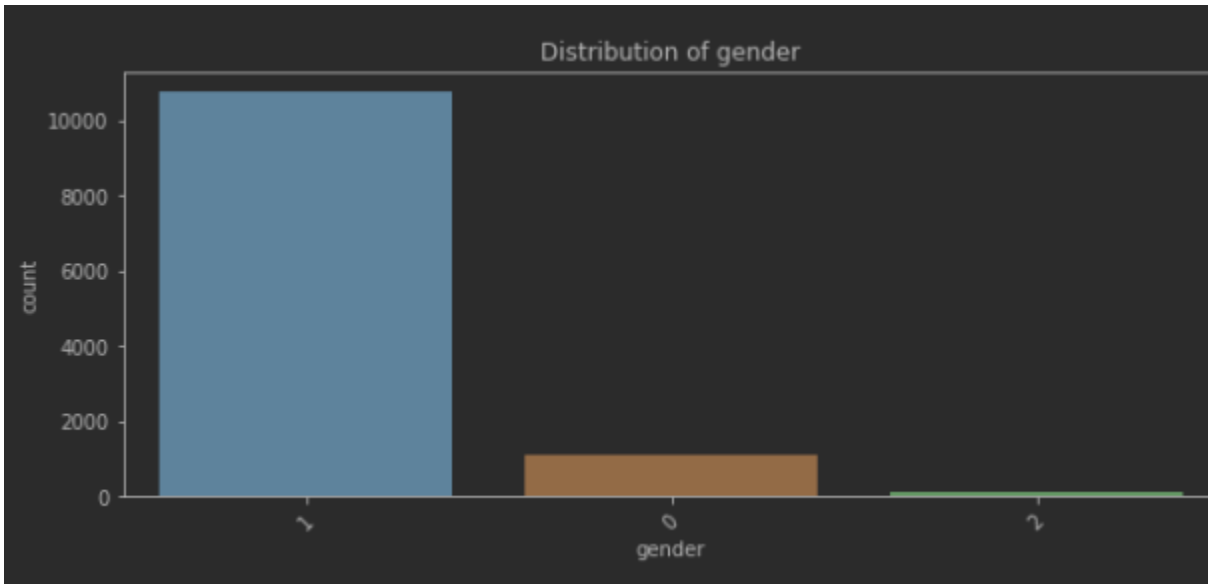
Visualisation

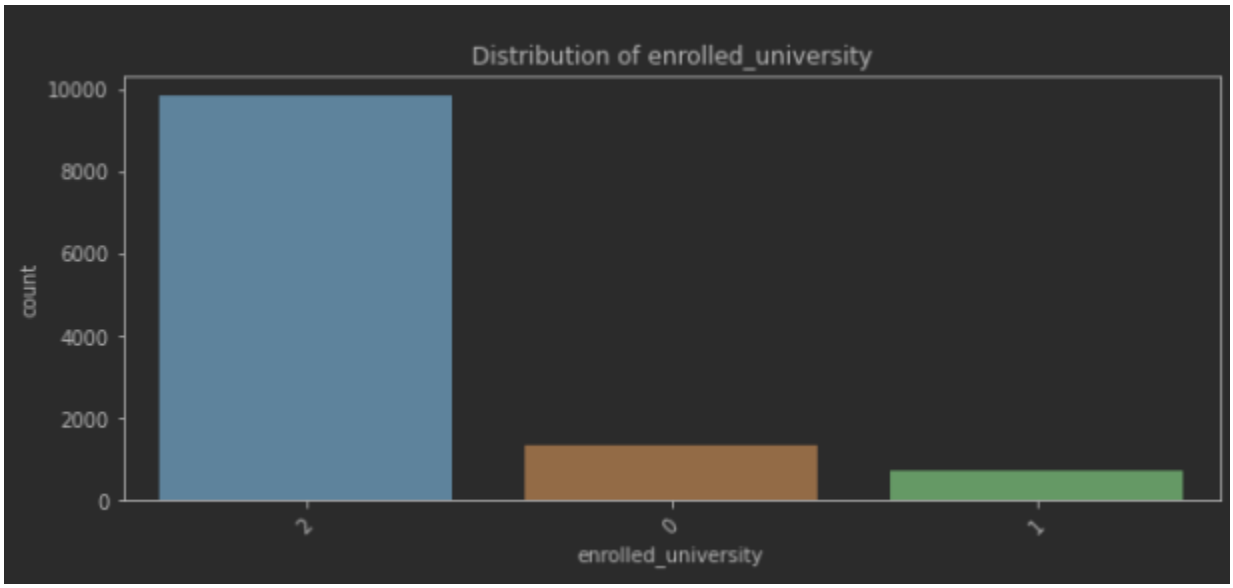
For the numerical columns (city_development_index and training_hours), histograms were plotted to get the distribution. Matplotlib library was used in the visualization.

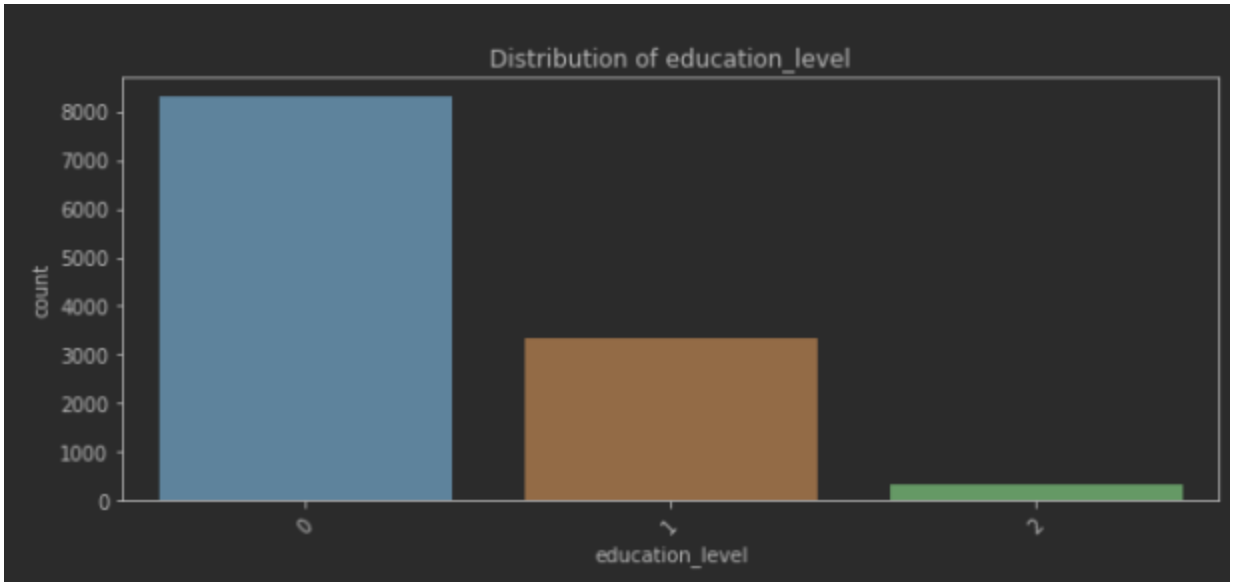


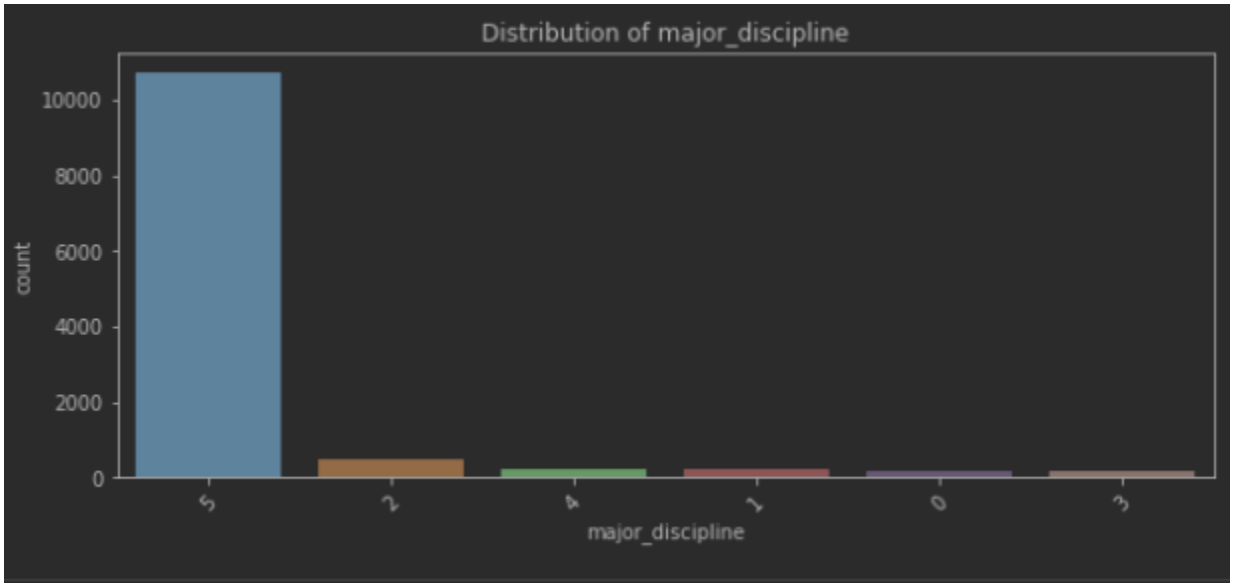


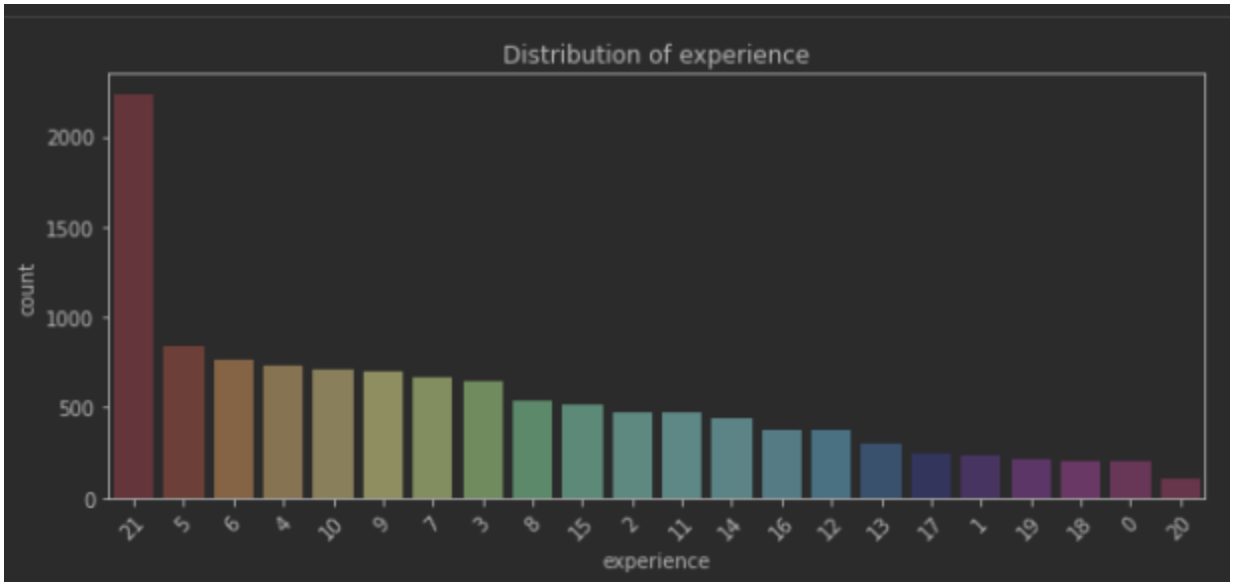
For the categorical data, bar charts were used to visualise

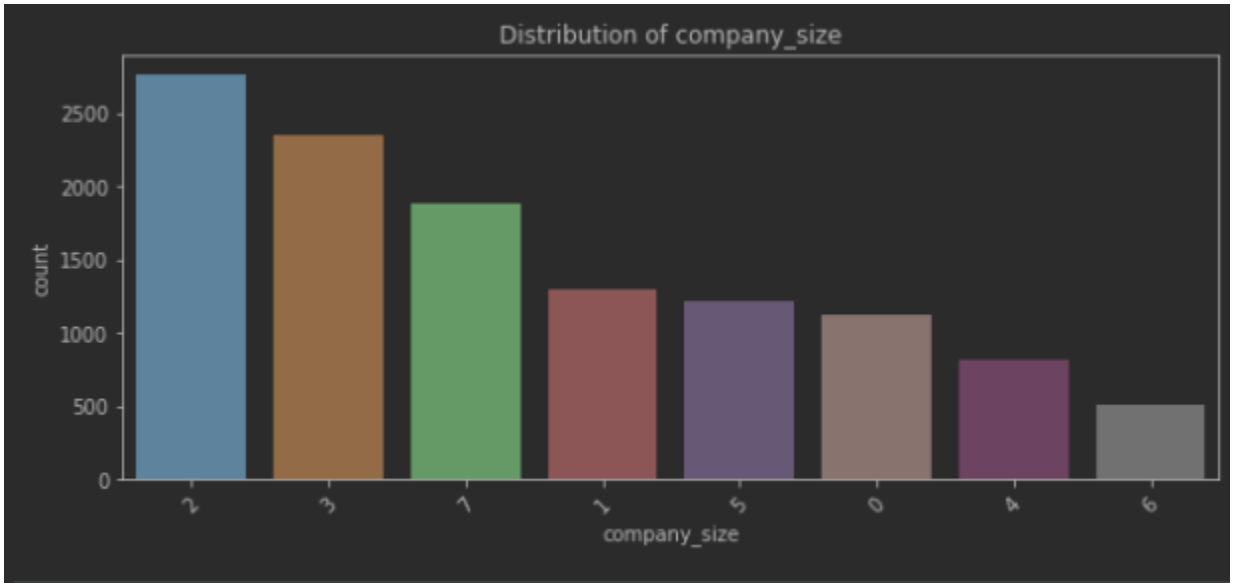


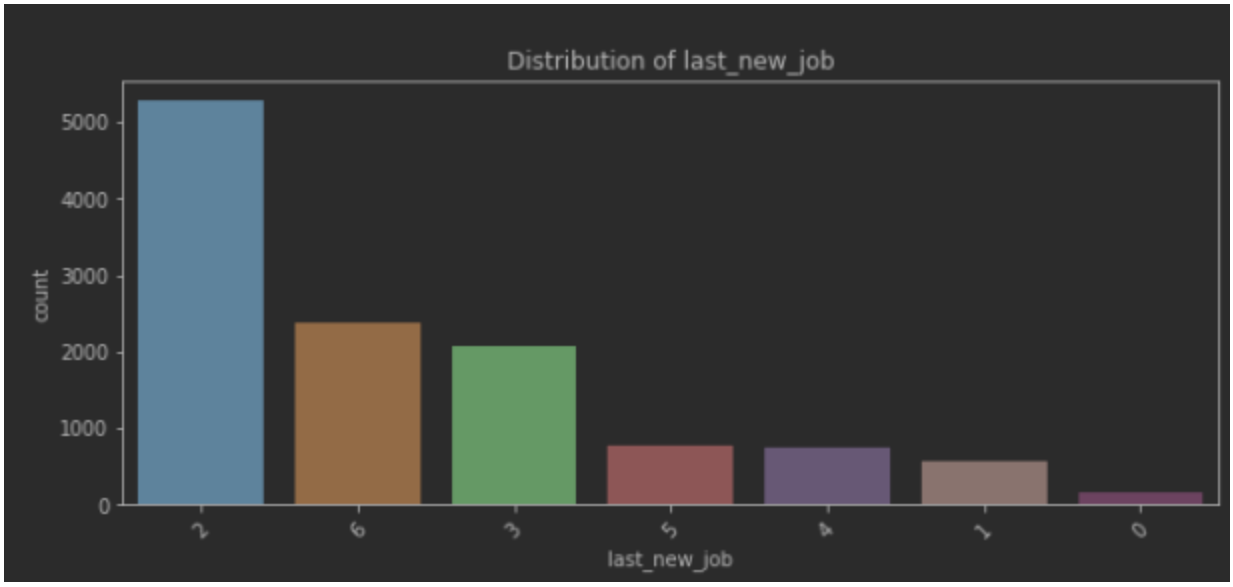


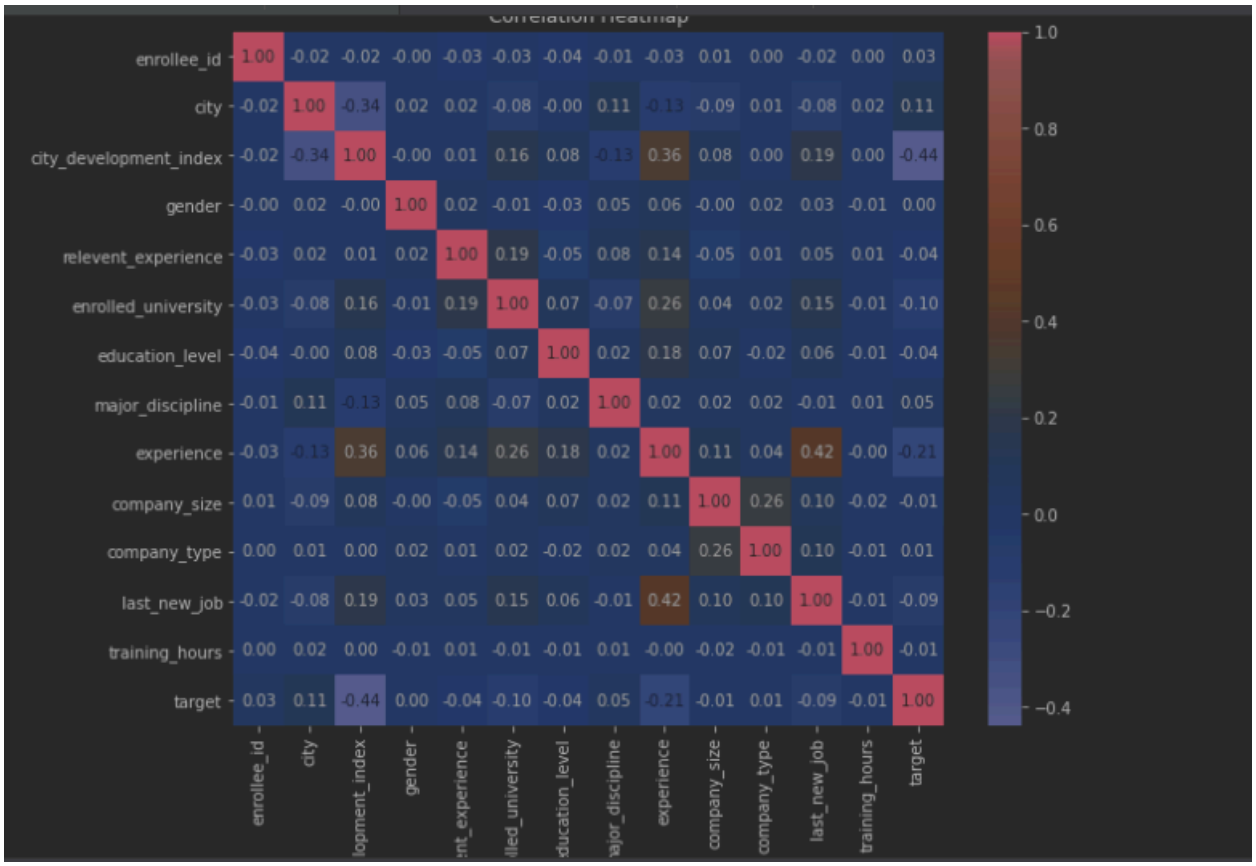

Distribution of education_level
count
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
education_level


Distribution of major_discipline
count
10000
8000
6000
4000
2000
0
5
2
4
1
0
3
major_discipline

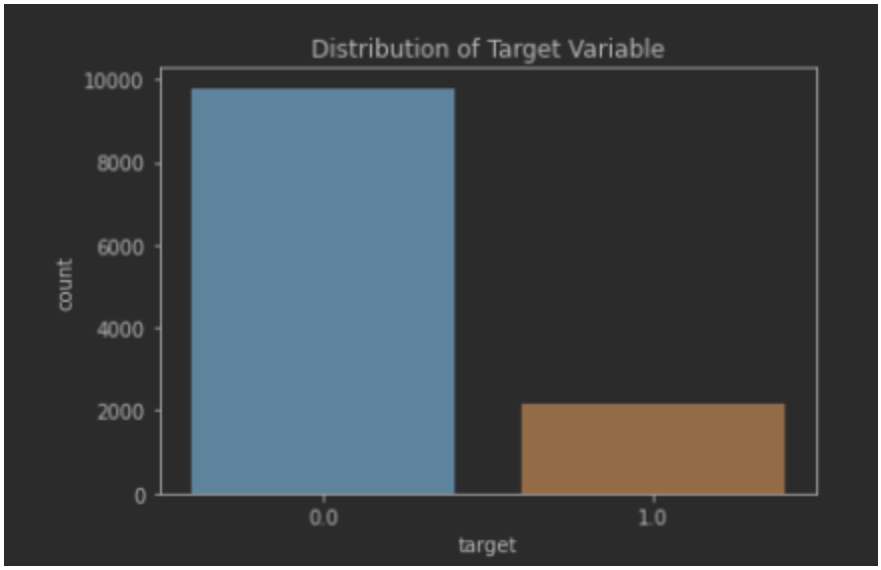

Distribution of experience
count
2000
1500
1000
500
0
21 5 6 4 10 9 7 3 8 15 2 11 14 16 12 13 17 1 19 18 0 20
experience


Distribution of company_size
count
2500
2000
1500
1000
500
0
2 3 7 1 5 0 4 6
company_size



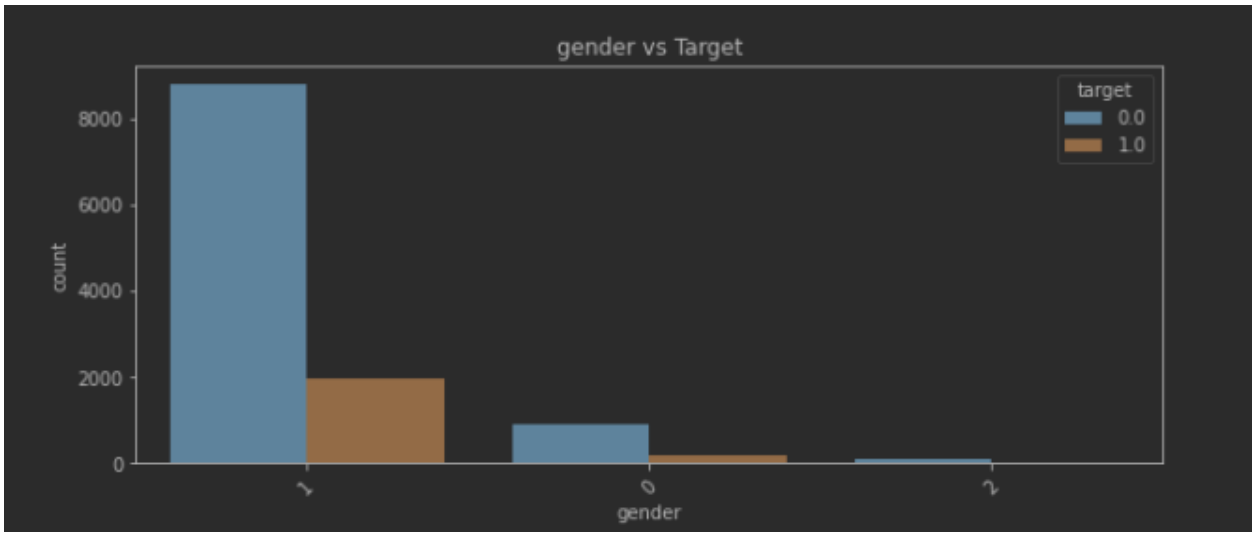


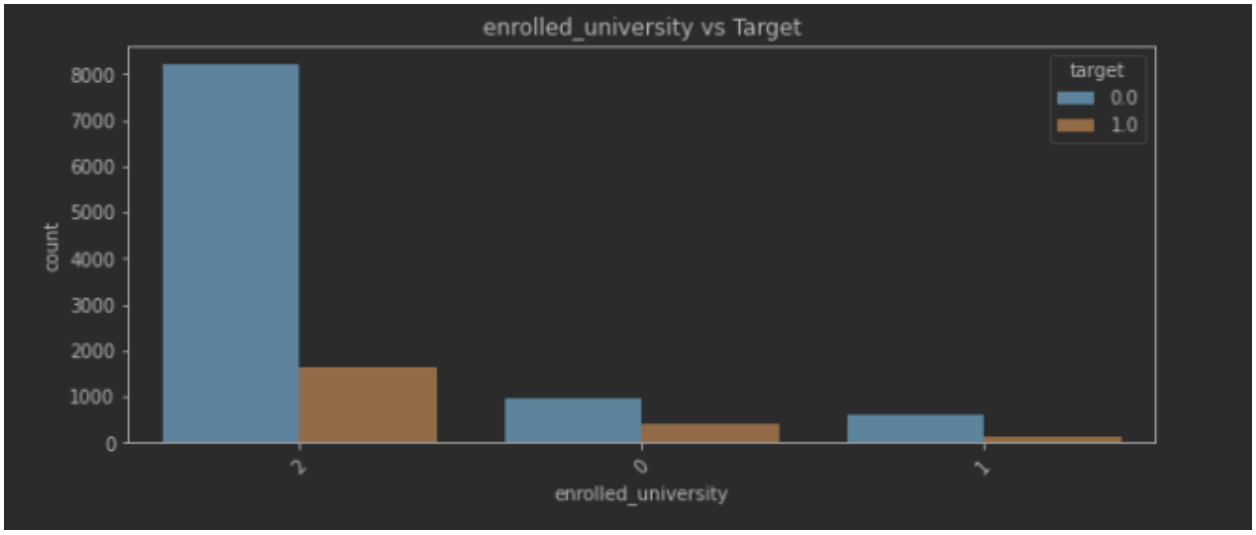
The visualisation above shows the correlation matrix, which is used in determining how different features relate with each other.

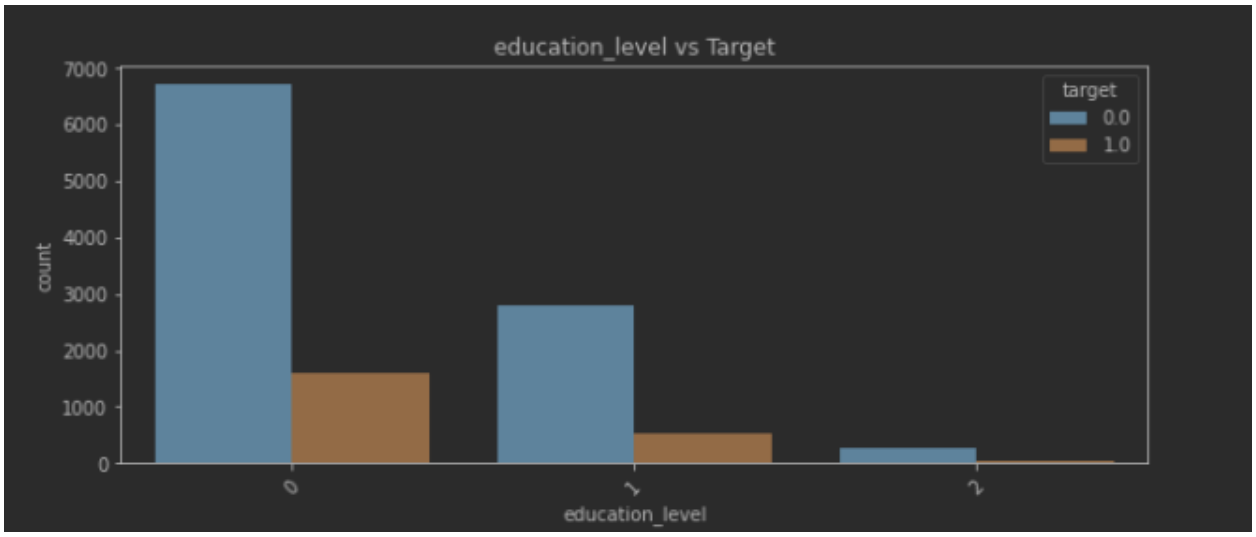


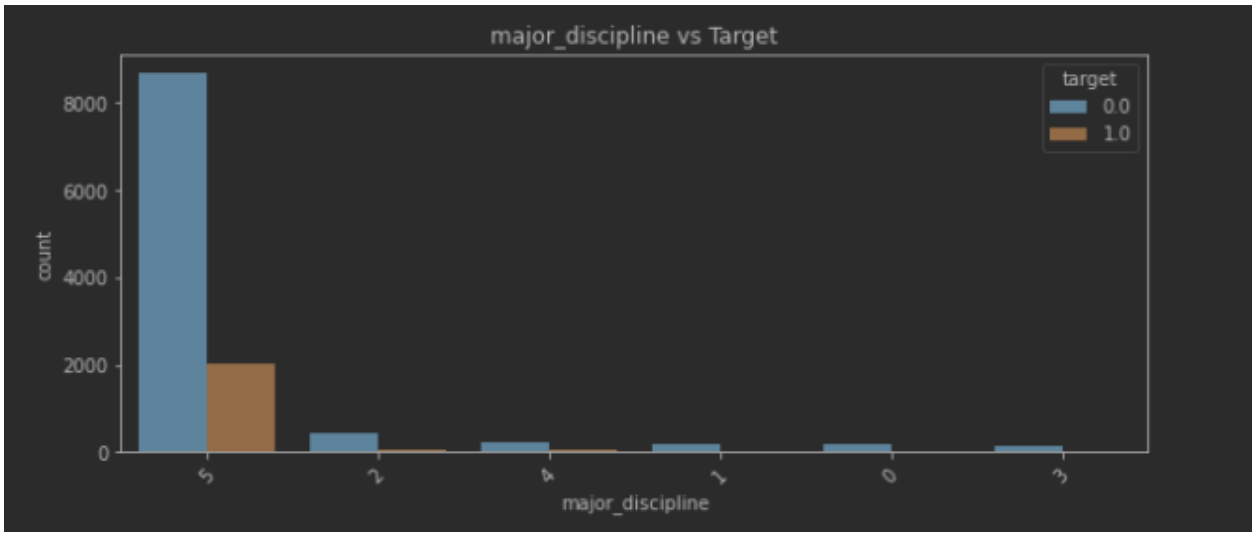
The visualisation above shows the distribution of the target variable.

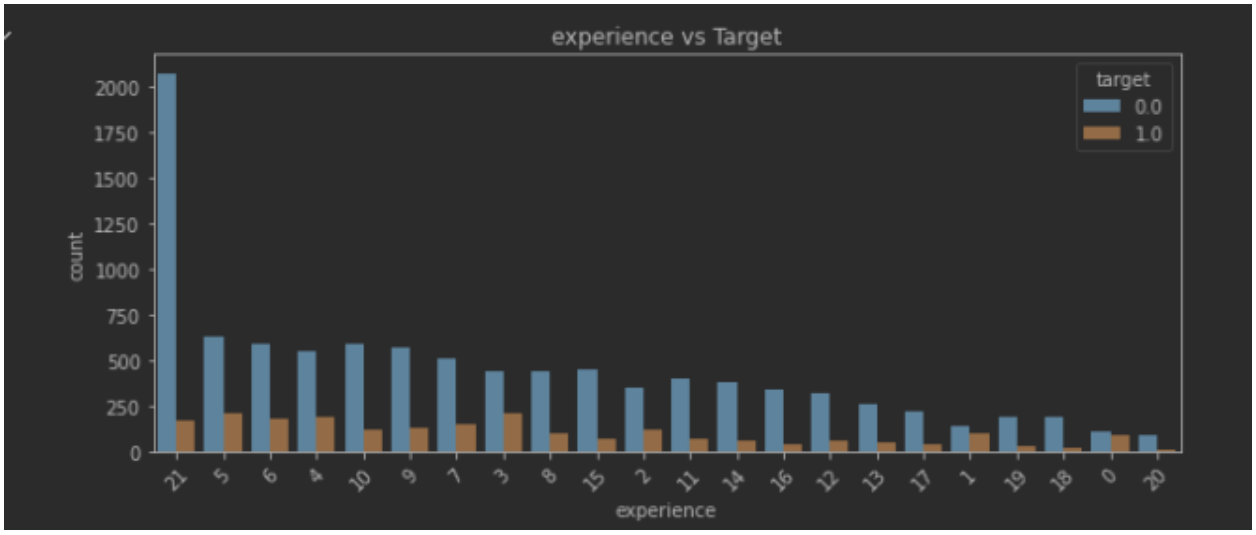
The visualisations below show the relationship between the target variable and other features

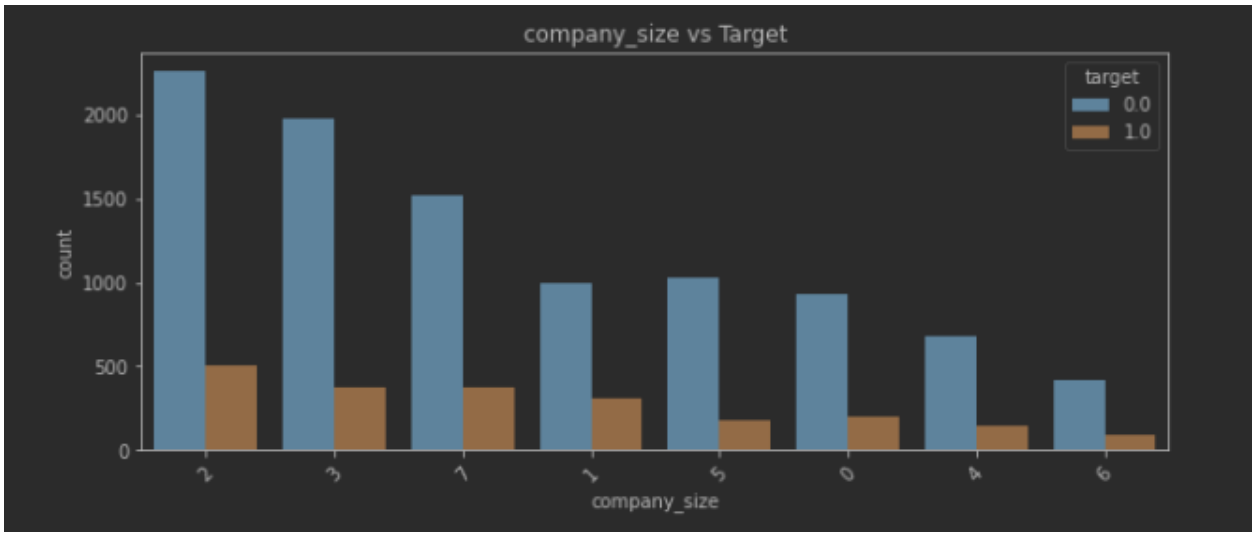


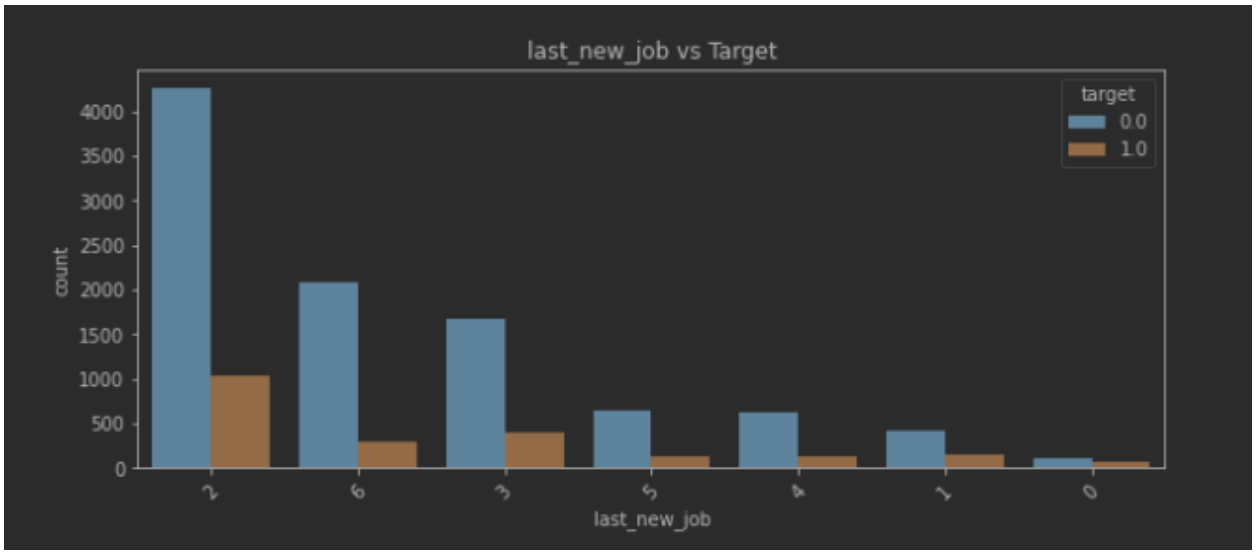

enrolled_university vs Target
target
0.0
1.0
count
8000
7000
6000
5000
4000
3000
2000
1000
0
2
0
1
enrolled_university

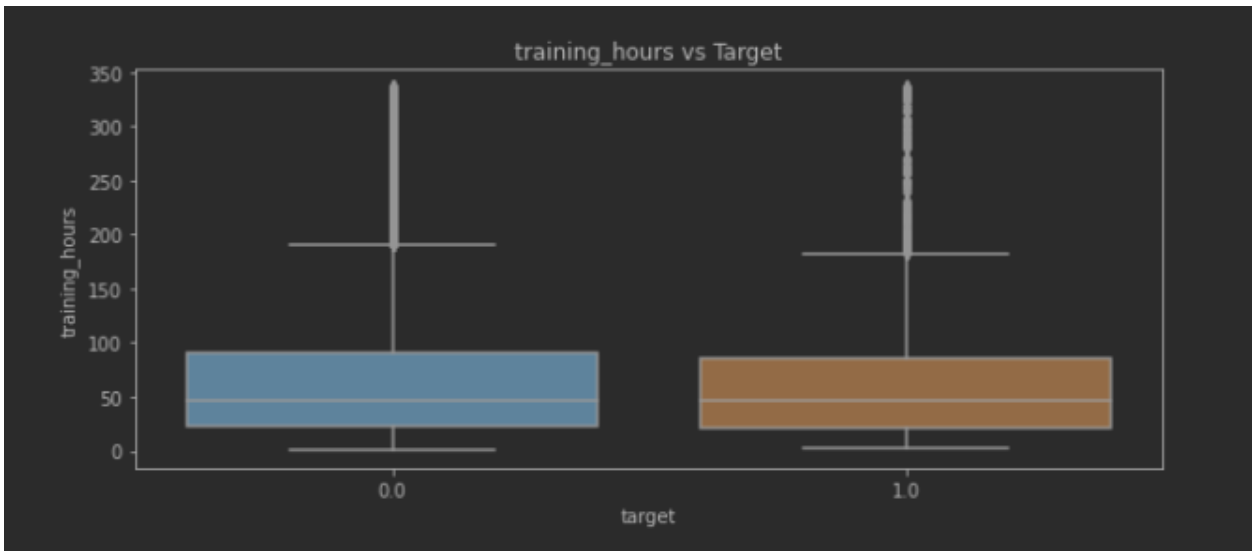

education_level vs Target
target
0.0
1.0
count
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
education_level


major_discipline vs Target
target
0.0
1.0
count
8000
6000
4000
2000
0
5
2
4
1
0
3
major_discipline


experience vs Target
target
0.0
1.0
count
2000
1750
1500
1250
1000
750
500
250
0
21
5
6
4
10
9
7
3
8
15
2
11
14
16
12
13
17
1
19
18
0
20
experience


company_size vs Target
target
0.0
1.0
count
2000
1500
1000
500
0
2
3
7
1
5
0
4
6
company_size


last_new_job vs Target
target
0.0
1.0
count
4000
3500
3000
2500
2000
1500
1000
500
0
2
6
3
5
4
1
0
last_new_job



Recommendations

The company should focus on the candidates that have higher training hours because they have a higher retention rate as compared to the others

The company should tailor training programs based on an employee's major discipline and education level to improve on retention.