Dataset

We will work on the adult dataset provided by the UCI1 website, which was collected by the United States Census Bureau (USCB), responsible for collecting demographic and economic information about individuals.

[http://archive.ics.uci.edu/dataset/2/adult](http://archive.ics.uci.edu/dataset/2/adult1)

The database contains about 49,000 samples and includes 14 characteristics to describe individuals, which we will mention in the following table:

|  |  |  |
| --- | --- | --- |
| Type | Feature | Id |
| numeric | Age | 1 |
| categorical | workclass | 2 |
| numeric | fnlwgt | 3 |
| categorical | education | 4 |
| numeric | education\_num | 5 |
| categorical | martial\_status | 6 |
| categorical | occupation | 7 |
| categorical | relationship | 8 |
| categorical | race | 9 |
| categorical | sex | 10 |
| numeric | capital\_gain | 11 |
| numeric | capital\_loss | 12 |
| numeric | hours\_per\_week | 13 |
| categorical | native\_country | 14 |

Preprocessing

First, Removing duplicate values from the database, about 24 values. Then calculating the missing values as a percentage and replacing them with the most frequent value (mode) of the corresponding attribute.

The features are mainly divided into six numerical and eight categorical features.

According to the numerical features, trying to explore and study the linear relationship and showing a heat map of the linear correlation matrix to help identify the relation of every pair of features in order to eliminate one of them.

Performing a standardization process (z-score normalization) for numerical features due to their importance for algorithms such as logistic regression.

Transforming numerical features to categorical by digitizing, but it turned out to be unhelpful because the data is not normally distributed and its order is important.

For the categorical features, representing them as vectors with one-hot encoding, removing the low frequencies for the categories (which contain zeros at a high rate of more than 99 percent), then deleting the first vector of each feature, which can be represented in other categories as zeros.

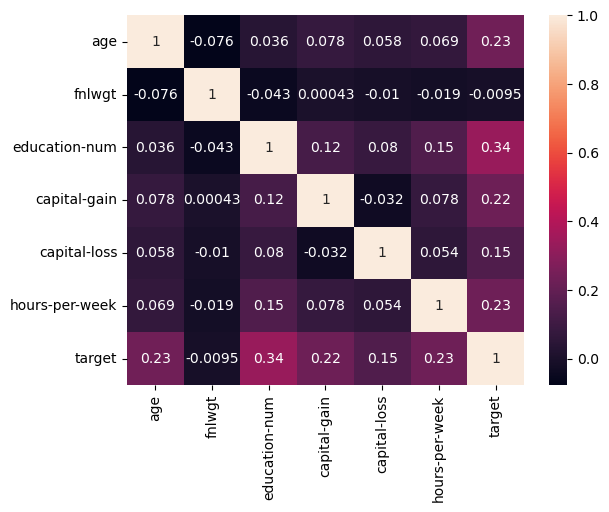


figure 1

Data Exploration

Based on the heat map in Figure 1, we studied the relationship between pairs of features in order to retain the important features in the training process. We noticed that the relationship between the pairs is generally weak, so we cannot get rid of one of the pairs and be satisfied with the other. Also from Figure 1, we notice that there is a weak correlation between the target and the fnlwgt attribute, which will be suggested to remove it after ensuring that it will not affect the accuracy of the training process.

For categorical features, we removed several features, such as the education feature, which has an equivalent attribute, education\_num, which is numerical, and other features such as native\_country, race, and sex, based on the experimentation process.

Testing

Splitting the dataset into 80% for fine tuning, and 20% for final testing.

For the baseline model, which is a dummy classifier model based on the mode, to determine the minimum expected performance, we obtained an accuracy of 75%. We can explain the result easily by the imbalance of the data.

Because of the imbalance of the dataset, we need another criterion beside accuracy to evaluate each model, so we adopted the area under curve (ROC) in the optimization process to reach the best hyperparameters of models.

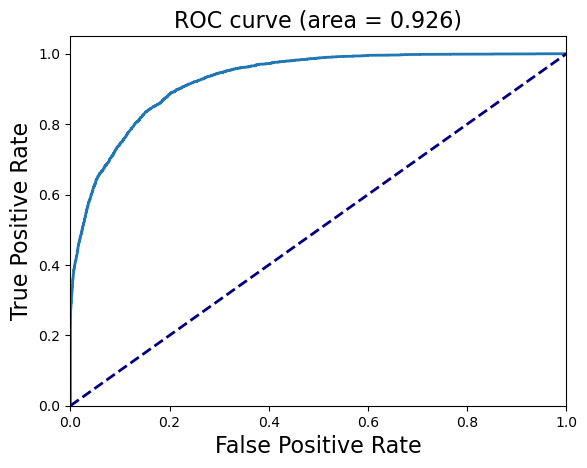
In the following table, the accuracy of each model is shown according to the training dataset.

|  |  |  |
| --- | --- | --- |
| **ROC** | **Accuracy** | **Model** |
| - | 75% | Dummy classifier |
| 0.911 | 85% | Logistic regression |
| 0.905 | 85% | K-nearest neighbor |
| 0.896 | 85% | Decision tree |
| 0.926 | 86.7% | AdaBoost |
| 0.929 | 86.7% | Gradient Boosting |

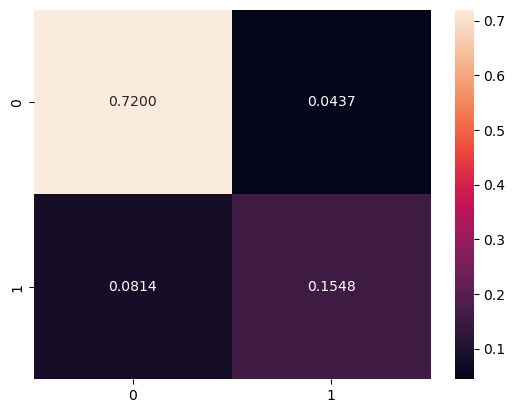
The best model that achieved the highest accuracy and highest degree of generalization, which is GBoost, with a maximum depth equals to 4 and a number of estimators equal to 250, the table below shows the results were obtained for the best model based on testing dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **AUC** | **Precision** | **Recall** | **Accuracy** |
| 0.9261942966661167 | 0.779771110423755 | 0.6554862194487779 | 0.8748848350838401 |

Here are two figures, the Rock curve and the confusion matrix. The confusion matrix represents horizontally the actual distribution, and vertically what was classified by the model, and the darkness of the secondary diameter indicates that the model made good prediction on the test data.



figure



figure

Conclusion

In this research, we tested a number of automatic learning models and documented the accuracy of each algorithm in the table above. In conclusion, we would like to clarify that the dataset is specific to individuals in the United States of America. So the results are not generally accurate due to the presence of several standards and factors related to the income of individuals depending on the geography.