Finding donors using Supervised Learning

Dataset: Adult income dataset

- The adult income dataset is a multivariate dataset.
- Total number of records: 45222 (After removing null valued records)
- Total attributes = 14

Attribute	Type of attribute
Age	Discrete
Work class	Nominal
Education level	Discrete
Education number	Discrete
Marital status	Nominal
Occupation	Discrete
Relationship	Nominal
Race	Nominal
Sex	Nominal
Capital gain	Continuous
Capital loss	continuous
Work hours per week	continuous
Native country	Nominal
Income (Target variable)	Asymmetric Binary

- Data splitting:
  1. Training data – 80% and
  2. Testing data – 20%
- Data splitting:
  1. Feature variables = 13
  2. Target variable = 1

Features

- For each person, 13 features were considered before feature encoding.
- After one hot encoding, total features encoded are 103.

Model and Algorithms:

- The Logistic regression algorithm is a predictive analysis algorithm based on probability. It is used for classification problems.
- The hypothesis of logistic regression tends to limit the cost function between 0 and 1.
- The K-Nearest Neighbors is another machine learning algorithm used in solving this problem of predicting donors. It can be useful for both regression and classification problems.
- KNN model implementation is done simply in few steps as:
  i. Load the data and initialize the value of K.
  ii. For getting predicted class, iterate from 1 to the total number of training data points.
  iii. Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other parameters that can be used are Chebyshev, cosine, etc.
  iv. Sort the calculated distances in ascending order based on distance values.
  v. Get top K rows from the sorted array.
  vi. Get the most frequent class of these rows and return the predicted class.
- Applying both models will give a chance of choosing the best fit model for the source data.

Approach

1. Load dataset from the source – Adult income dataset.
2. Data preparation and visualization.
3. Feature encoding and normalization.
4. Splitting data into training data and testing data.
5. Naive predictor performance checking.
6. Dividing batches and training supervised learning models using two supervised algorithms/classifiers as:
   i. Logistic Regression
   ii. K-Nearest Neighbors
7. Predicting the testing data using the above classifiers.
8. Drawing the confusion matrix and measuring the accuracy of models.

Result:

Based on the source data and machine learning algorithms, successfully training and testing model for finding donors using supervised learning has done with resulting accuracy scores for both the models. And “Logistic regression” has the best accuracy score when compared to the K-Nearest Neighbor algorithm accordingly to the data splitting.

References

1. Dataset: https://archive.ics.uci.edu/ml/datasets/Adult
2. Concept research: https://www.donorsearch.net/finding-new-donors/
3. Logistic regression overview: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
4. Sklearn linear model - Logistic regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
5. K-Nearest Neighbors algorithm study: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
6. Understanding variables: https://socratic.org/questions/is-age-continuous-or-discrete-data#:~:text=Answer%3A%20Continuous%20if%20looking%20for,any%20value%20within%20the%20range.
7. Attributes and classification: https://www.geeksforgeeks.org/understanding-data-attribute-types-qualitative-and-quantitative/