6/2/2019

Assignment 2: Data Modelling and Presentation

Practical Data Science | COSC2670

SARTHAK SIRARI | S3766477
MASTER OF DATA SCIENCE | MC267
COMPUTER SCIENCE & IT, SCHOOL OF SCIENCE
RMIT UNIVERSITY
S3766477@STUDENT.RMIT.EDU.AU

INDEX

EXECUTIVE SUMMARY

The aim of this report is to predict whether a person has an income of over 50K a year or not with the given set of demographic information. It also checks which algorithm predicts results with better accuracy. Further, to check relationships between the various factors in the data set.

We chose a data set from the UCI Machine Learning Repository containing the demographics of a set of population from around the world including factors like age, gender, race, education, income, etc. and performed K-Nearest Neighbour and Decision Tree algorithm for predictions. Furthermore, we checked for relationships between the factors within the data set.

The accuracy rate for K-Nearest Neighbour algorithm was better for predicting whether a person has an income of over 50K a year or not than the Decision Tree algorithm with default values.

The report concludes that we must K-Nearest Neighbour algorithm with default values for predicting whether a person has an income of over 50K a year or not as it has the better accuracy rate. Moreover, there are significant relationships present between the factors present in the data set.

However, it is recommended that we must include more factors into account such as years of experience, to have a better accuracy in predicting results.

INTRODUCTION

We know that there are several factors which depicts the income of an individual, such as age, education, gender, native country, etc. Even though there can be huge differences between the incomes of two individuals with the same demographic conditions at the real-world scenario, we wanted to have a general look at a small subset of population from around the world.

We wanted to check if there is a way, we could use some demographic information and determine if we could predict the income of an individual based to these factors. Further, we wanted to find any relationships between these factors involved for this prediction of income of individuals.

This report will discuss if we could predict whether a person has an income of over 50K a year or not, by using simple demographic information.

METHODOLOGY

1. Data Collection

To conduct our study, we used an open data set available freely at the UCI Machine Learning Repository, named Adult Data Set, present under the link (http://archive.ics.uci.edu/ml/datasets/Adult).

This data set contains 14 attributes more than 32,000 observations and was made public by Ronny Kohavi and Barry Becker of the Silicon Graphics in the year 1996. This data set contains a set of demographic information which may decide whether an individual makes 50K a year or not.

2. Importing Libraries and Data Loading

Firstly, we made all the necessary imports. Secondly, we read the provided CSV/DATA file by using *read_csv()* function present in the pandas library. We verify if the data is loaded into the memory by printing it into console by using commands like *print(adult)* and *adult.head()*.

3. Data Cleaning

Once our data is loaded, we must check for possible errors in our data as discussed further.

Following issues with data were removed while preparing data: -

1. Missing values
2. Impossible Values
3. Typos
4. Case Mismatch
5. Extra Whitespace

4. Encoding Labels

We used LabelEncoder() to give numbers to categorical values to further prepare our data set

5. Diving Data into Factors

We divided the data into input and output sets by using iloc() command.

6. Feature Selection

We used F-Score based Select K Best feature selection method by using the SelectKBest() method and selected 10 best columns for further study.

7. Predicting results using K-Nearest Neighbour

We divided the data into training and test data and performed the K-Nearest Neighbour algorithm as following: -

1. 50% for training and 50% for testing
2. 60% for training and 40% for testing
3. 80% for training and 20% for testing

We used the default values as they gave good result with an accurate of more than 80%.

8. Predicting results using Decision Tree

We divided the data into training and test data and performed the K-Nearest Neighbour algorithm as following: -

1. 50% for training and 50% for testing
2. 60% for training and 40% for testing
3. 80% for training and 20% for testing

We used the default values as they gave good result with an accurate of more than 80%.

RESULTS

1. Results using K-Nearest Neighbour

1.1. 50% for training and 50% for testing

Following are the result of the test: -

```
Confusion Matrix:
[[10067  1212]
 [ 1494  2308]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.89      0.88     11279
           1       0.66      0.61      0.63      3802

   micro avg       0.82      0.82      0.82     15081
   macro avg       0.76      0.75      0.76     15081
weighted avg       0.82      0.82      0.82     15081

Accuracy of the K-Nearest Neighbour Model is:  0.8205689277899344
```

1.2. 60% for training and 40% for testing

Following are the result of the test: -

```
Confusion Matrix:
[[8090  918]
 [1220 1837]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.90      0.88      9008
           1       0.67      0.60      0.63      3057

   micro avg       0.82      0.82      0.82     12065
   macro avg       0.77      0.75      0.76     12065
weighted avg       0.82      0.82      0.82     12065

Accuracy of the K-Nearest Neighbour Model is:  0.8227932034811438
```

1.3. 80% for training and 20% for testing

Following are the result of the test: -

```
Confusion Matrix:
[[4063  445]
 [ 608  917]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.90      0.89      4508
           1       0.67      0.60      0.64      1525

   micro avg       0.83      0.83      0.83      6033
   macro avg       0.77      0.75      0.76      6033
weighted avg       0.82      0.83      0.82      6033

Accuracy of the K-Nearest Neighbour Model is:  0.8254599701640974
```

2. Results using Decision Tree

2.1. 50% for training and 50% for testing

Following are the result of the test: -

```
Confusion Matrix:
[[9883 1396]
 [1507 2295]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.88      0.87     11279
           1       0.62      0.60      0.61      3802

   micro avg       0.81      0.81      0.81     15081
   macro avg       0.74      0.74      0.74     15081
weighted avg       0.81      0.81      0.81     15081

Accuracy of the Decision Tree Model is:  0.8075061335455208
```

2.2. 60% for training and 40% for testing

Following are the result of the test: -

```
Confusion Matrix:
[[7948 1060]
 [1226 1831]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.88      0.87      9008
           1       0.63      0.60      0.62      3057

   micro avg       0.81      0.81      0.81     12065
   macro avg       0.75      0.74      0.74     12065
weighted avg       0.81      0.81      0.81     12065

Accuracy of the Decision Tree Model is:  0.8105263157894737
```

2.3. 80% for training and 20% for testing

Following are the result of the test: -

```
Confusion Matrix:
[[4019  489]
 [ 621  904]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.89      0.88      4508
           1       0.65      0.59      0.62      1525

   micro avg       0.82      0.82      0.82      6033
   macro avg       0.76      0.74      0.75      6033
weighted avg       0.81      0.82      0.81      6033

Accuracy of the Decision Tree Model is:  0.8160119343610144
```

3. Hypothesis while comparing data

- **Income v/s Education**: Income increases with education level.
- **Income v/s Hours Per Week**: People who work more hours per week has better income.
- **Income v/s Gender**: Men earn more than women.
- **Income v/s Relationship**: Couples have more income than and Singles.
- **Income v/s Age**: People who earn greater than 50K are older than who earn less than 50K.
- **Capital Gain v/s Work Class**: People who are self-employed have better capital gains.
- **Capital Loss v/s Work Class**: Capital Losses are experienced mostly by people with Bachelors and High School grads.

- **Hours Per Week v/s Country**: Greek and Thai people work most number of hours per week.
- **Hours Per Week v/s Gender**: Men work more hours per week than Women.
- **Hours Per Week v/s Occupation**: Farmers and Fishermen work most hours per week.

DISCUSSION

We can observe that the accuracy rate for K-Nearest Neighbour algorithm is better than Decision Tree algorithm to predict whether a person has an income of over 50K a year as shown following for the default values: -

- 50% for training and 50% for testing

  K-Nearest Neighbour: **0.821**

  Decision Tree: **0.816**

- 60% for training and 40% for testing

  K-Nearest Neighbour: **0.825**

  Decision Tree: **0.811**

- 80% for training and 20% for testing

  K-Nearest Neighbour: **0.825**

  Decision Tree: **0.812**

CONCLUSION

The report concludes that we must K-Nearest Neighbour algorithm with default values for predicting whether a person has an income of over 50K a year or not as it has the better accuracy rate. Moreover, there are significant relationships present between the factors present in the data set.

Hence, K-Nearest Neighbour has a better accuracy rate for all given split of data with default values.

REFERENCES

- UCI MACHINE LEARNING LIBRARY, Adult Data Set (http://archive.ics.uci.edu/ml/datasets/Adult).