Machine-learning analysis for adult census income dataset

Article · April 2020

CITATIONS
0

READS
4,231

1 author:

Ahmed Ali
University of the Cumberlands
44 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Ahmed Ali

Ph.D. Student

University of Cumberlands

Adult Census Income Dataset Analysis

Adult census income Dataset

The dataset extracted from Adult Census Income in 1994 by Ronny Kohavi and Barry Becker, the dataset includes 15 variables. One predication goal is to determine if a person makes over $50K a year based on the biography and background. Figure 1 and 2 shows the dataset variables.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356	40	United-States	<=50K
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356	40	United-States	<=50K
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K

Figure 1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education.num   32561 non-null  int64
 5   marital.status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital.gain    32561 non-null  int64
 11  capital.loss    32561 non-null  int64
 12  hours.per.week  32561 non-null  int64
 13  native.country  32561 non-null  object
 14  income          32561 non-null  object
```

Figure 2

Data processing

Figure 3 shows the first step before performing the data exploratory is performing data cleaning by removing spaces and unwanted characters and removing the duplicates rows. Next, we check for null. Next, counting and removing rows that occupation is '?'. Also, delete "fnlwgt" feature because it has no effect for predicting the income, whether exceed 50k.

Also, for validation, we are splitting the datasets into training sets and test sets. The data processing also include encoding categorical data categorical variables include : [workclass:1,education:2, marital-status:4, occupation:5, relationship:6, race:7, sex:8, native-country:12]

Exploratory data analysis

Figure 7 shows the exploratory data analysis of the distribution of the age of people across the dataset. The plot shows that the most significant numbers of participants ages of the dataset are between 30 and 40, while a small number of people less than 20 and over 60, the minimum age is 17. Figure 9 shows the distribution of different features of the dataset; the plot shows different features such as the hours per week is 40, the ages, capital loss, education, and so on.

Figure 3 Remove the spaces in names

```
print(df_adult_eda.shape)

# Dropping the duplicate Rows
df_adult_eda = df_adult_eda.drop_duplicates(keep = 'first')
df_adult_eda.shape
```

```
(32537, 15)
(32537, 15)
```

Figure 4 Remove the duplicates in rows

```
# Checking the null values in the columns
df_adult_eda.isnull().sum(axis = 0)
```

```
age               0
workclass         0
fnlwgt            0
education         0
education.num     0
marital.status    0
occupation        0
relationship      0
race              0
sex               0
capital.gain      0
capital.loss      0
hours.per.week    0
native.country    0
income            0
dtype: int64
```

Figure 5 counts the null in all columns

```
# Dropping the rows whose occupation is '?'
df_adult_eda = df_adult_eda[df_adult_eda.occupation != '?']

df_adult_eda['occupation'].value_counts()
```

```
Prof-specialty       4136
Craft-repair         4094
Exec-managerial      4065
Adm-clerical         3768
Sales                3650
Other-service        3291
Machine-op-inspct    2000
Transport-moving     1597
Handlers-cleaners    1369
Farming-fishing       992
Tech-support          927
Protective-serv       649
Priv-house-serv       147
Armed-Forces            9
Name: occupation, dtype: int64
```

Figure 6 after removing occupation is '?'

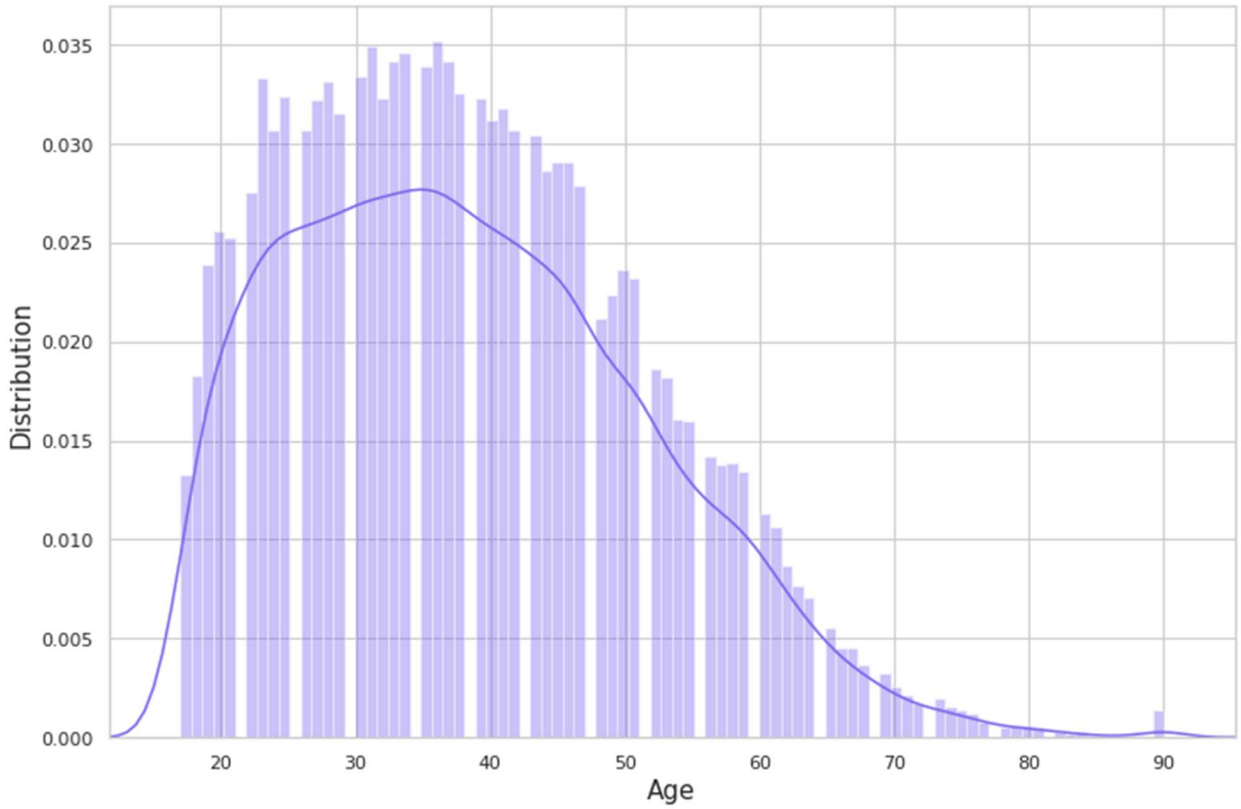

Figure 7 Distribution of age of people across the dataset

Figure 8 shows the heatmap shows the correlation between the different variables; the map shows a correlation in the ages and number of hours also, the education and income have tied correlation, as the income is increasing by the level of education. Figure 10 shows the hours per week, according to the education level of the person. The graph shows that the highest working hours were for the highest education. The people who have doctorate education have about 48 working hours, while for those who have only school education have fewer working hours per week. For the bachelor's group, the work hours have fewer working hours than the group with a master's degree. The number of hours can be used as an indicator for more income within the same education group; for

example, the higher education, which includes bachelor's, master, and doctorate, the figure shows the people with doctorate have working hours more than the bachelors.

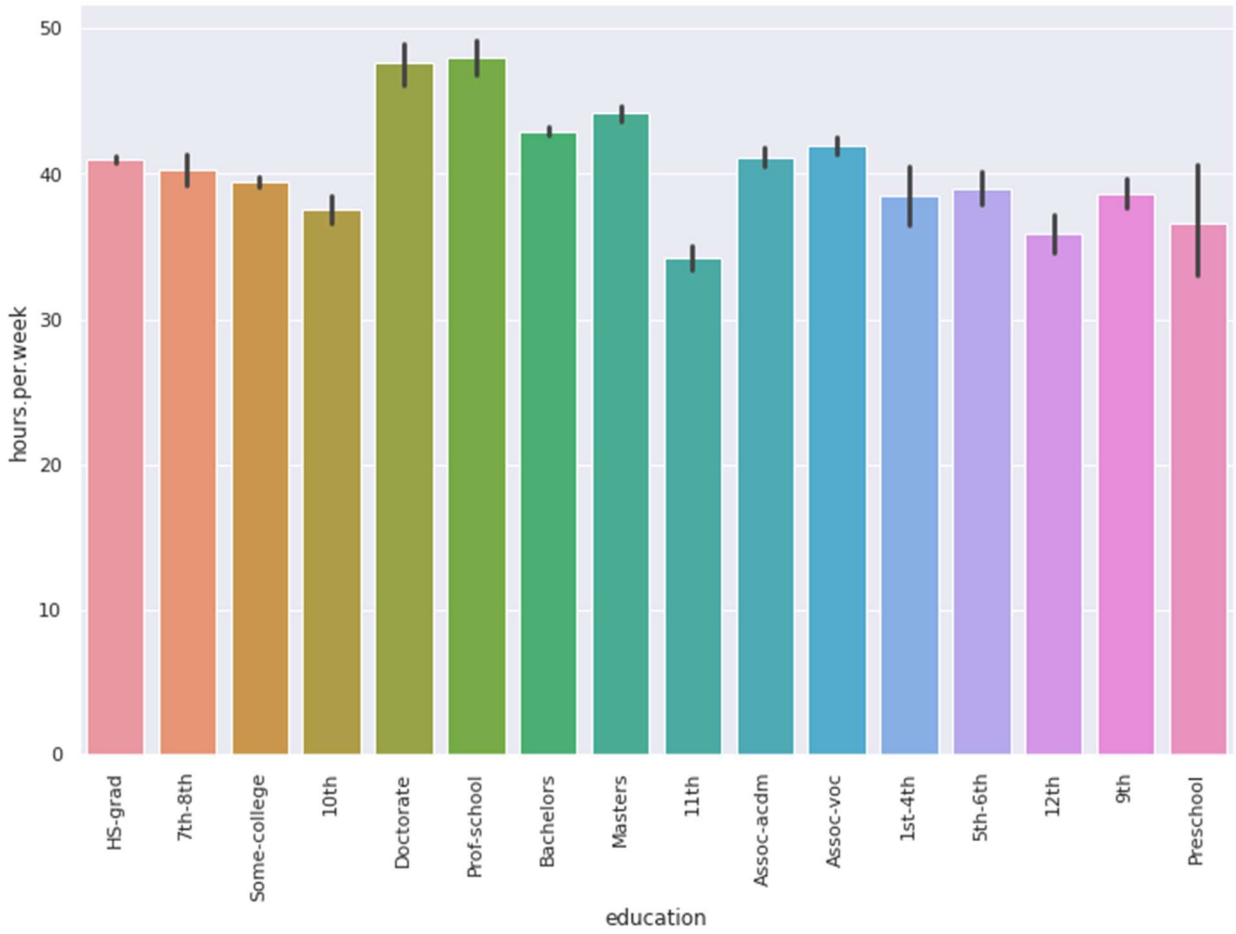

Figure 10 the hours per week according to the education of the person

Machine Learning – Decision Tree

Building the decision tree classification model, we used the Sklearn library for building the model. Then for evaluating the model, we compare the accuracy of models with different parameters.

Furthermore, it is vital to construct the confusion matrix that determines model sensitivity and specificity is given below, along with the ROC curve for the random oversampling model.

Table 1 shows the precision and recalls of the data model for those who have almost 50K. The precision is about 95%, while for greater than 50K, the precision is 58%. The average recall is about 83%.

Choosing the classifier three whose max_leaf_nodes is set to 16 having fewer branches than classifier 4, it can be easier for visualizing the model result and analyzing the result. Besides, its model accuracy is 0.846922 close to classifier four which is the best classifier among the 22 classifiers

classifier model	precision	recall	f1-score	support
at_most_50K	0.95	0.88	0.91	5391
greater_than_50K	0.58	0.80	0.67	1122
macro avg	0.77	0.84	0.79	6513
weighted avg	0.89	0.86	0.87	6513

```
classifier1 = DecisionTreeClassifier()
classifier1.fit(x_train, y_train)
y_predict1_test=classifier1.predict(x_test)
y_predict1_train=classifier1.predict(x_train)
classifier1
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

Figure 11 building the decision Tree model using Sklearn library

```
Classifier1(default) Accuracy: 0.817442
Classifier2(max_leaf_nodes=8) Accuracy: 0.844772
Classifier3(max_leaf_nodes=16) Accuracy: 0.851067
Classifier4(max_leaf_nodes=32) Accuracy: 0.860279
Classifier5(max_leaf_nodes=64) Accuracy: 0.864118
Classifier6(max_leaf_nodes=128) Accuracy: 0.861815
Classifier7(min_impurity_decrease=0.001) Accuracy: 0.856595
Classifier8(min_impurity_decrease=0.01) Accuracy: 0.844772
Classifier9(min_impurity_decrease=0.02) Accuracy: 0.818210
Classifier10(min_impurity_decrease=0.03) Accuracy: 0.762475
Classifier11(min_impurity_decrease=0.04) Accuracy: 0.762475
Classifier12(min_samples_leaf=40, min_samples_split=80) Accuracy: 0.859205
Classifier13(min_samples_leaf=80, min_samples_split=160) Accuracy: 0.861201
Classifier14(min_samples_leaf=160, min_samples_split=320) Accuracy: 0.853217
Classifier15(min_samples_leaf=320, min_samples_split=640) Accuracy: 0.853063
Classifier16(min_samples_leaf=640, min_samples_split=1280) Accuracy: 0.834792
Classifier17(max_leaf_nodes=16,min_impurity_decrease=0.001) Accuracy: 0.844772
Classifier18(max_leaf_nodes=16,min_impurity_decrease=0.01) Accuracy: 0.818210
Classifier19(max_leaf_nodes=16,min_impurity_decrease=0.02) Accuracy: 0.851067
Classifier20(max_leaf_nodes=16,min_samples_leaf=40, min_samples_split=80) Accuracy: 0.851067
Classifier21(max_leaf_nodes=16,min_samples_leaf=80, min_samples_split=160) Accuracy: 0.851067
Classifier22(max_leaf_nodes=16,min_samples_leaf=160, min_samples_split=320) Accuracy: 0.853831
```

Figure 12 compares the accuracy of models with different parameters

We have chosen the classifier model number three, whose maximum leaf nodes set to 16, which have fewer branches than classifier 4; it can be more comfortable for visualizing the model result and analyzing the result. Besides, its model accuracy is 0.84 close to classifier four, which is the best classifier among the 22 classifiers.

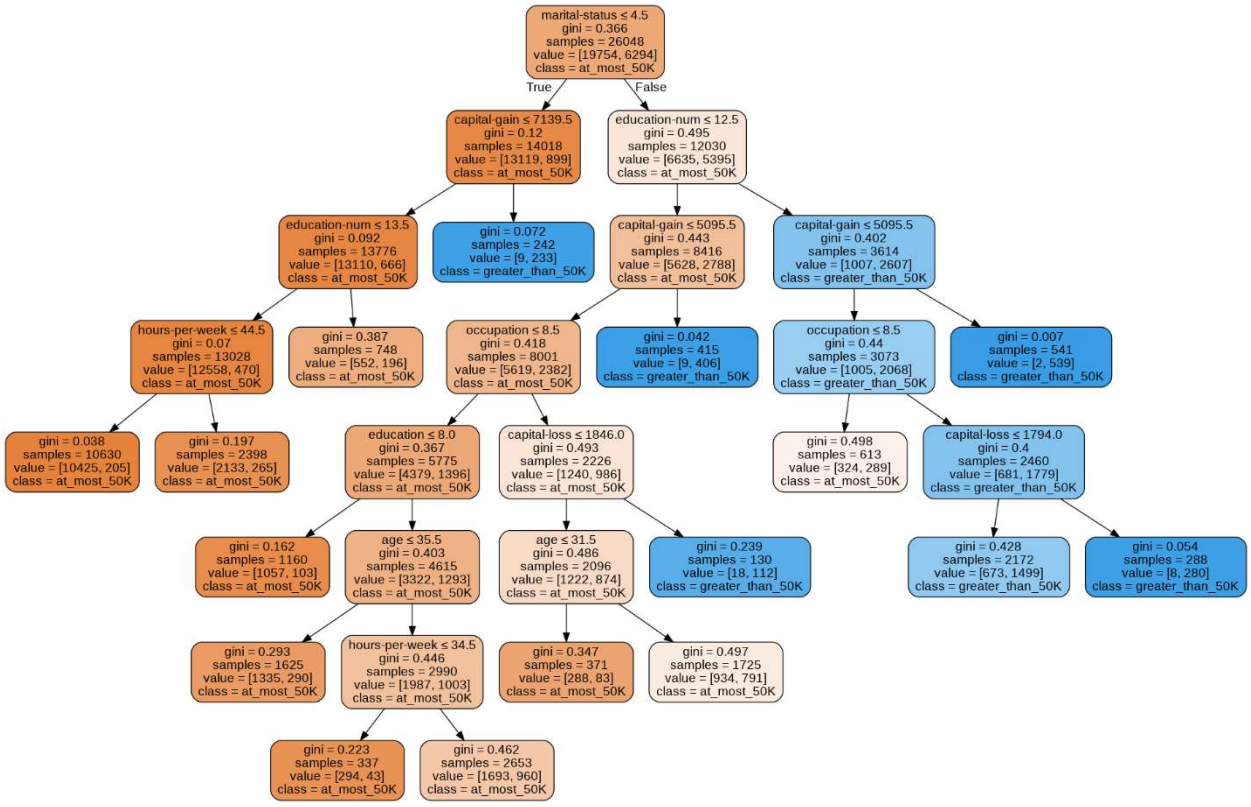

Figure 13 Decision tree model

Conclusion

This decision tree visualization graph shows that dataset variables that most have a direct impact are education level, capital income, capital loss, marital status, and occupation that will influence whether people make more than 50k per year. When people's marital status is married, and their education level is higher especially if they have a college

degree, furthermore, when they get more capital gain and more capital loss, then they will have a higher possibility that he or she will make more than 50k per year.

References

Mishra, S. (2019). Adult Census Income Predict income. Retrieved from https://www.kaggle.com/uciml/adult-census-income/discussion/126525

Kohavi, R.(2011). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. Retrieved from http://robotics.stanford.edu/~ronnyk/nbtree.pdf

Hong, M., Z.(2019), Income Prediction using CART Retrieved from https://www.kaggle.com/manzih/income-prediction-using-cart-accuracy-86-41

Lemon, C., Zelazo, C., Mulakaluri, K.. (2018), Predicting if income exceeds $50k per year based on 1994 US Census Data with Simple Classification Techniques Retrieved from http://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf