project

April 5, 2023

1 CSPB 3022 Project

1.1 Author: Adam Fowler

1.2 Project Topic

This project explores a dataset with various attributes of a person in order to determine if they are likely to get a job that pays $50k or more using binary classification. While this is primarily a classification exercise, if regression is applicable I would like to apply that as well.

The goal of this project is to be able to correctly classify these instances. My motivation is simply to learn the techniques involved.

1.3 Data

The dataset used is available from Kaggle. No information is given on how the data was gathered.

> Gal Shochat. (2022). Classification problem/ Yes or NO 50K salary, Version 1. Retrieved April 5, 2023 from https://www.kaggle.com/datasets/galshochat/classification-problem-yes-or-no-50k-salary.

[ ]:
```
import pandas as pd, numpy as np
```

[ ]:
```
df = pd.read_csv('adult.data');
print(f'{df.shape = }')
df.head()
```

```
df.shape = (32560, 15)
```

[ ]:

	39	State-gov	77516	Bachelors	13	Never-married	\
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	
1	38	Private	215646	HS-grad	9	Divorced	
2	53	Private	234721	11th	7	Married-civ-spouse	
3	28	Private	338409	Bachelors	13	Married-civ-spouse	
4	37	Private	284582	Masters	14	Married-civ-spouse	

	Adm-clerical	Not-in-family	White	Male	2174	0	40	\
0	Exec-managerial	Husband	White	Male	0	0	13	
1	Handlers-cleaners	Not-in-family	White	Male	0	0	40	

```
2   Handlers-cleaners         Husband   Black     Male      0   0   40
3      Prof-specialty            Wife   Black   Female      0   0   40
4     Exec-managerial            Wife   White   Female      0   0   40

    United-States   <=50K
0   United-States   <=50K
1   United-States   <=50K
2   United-States   <=50K
3            Cuba   <=50K
4   United-States   <=50K
```

The data is tabulated with 32560 samples and 15 features.

Feature	Type
Age	Integer
Workclass	Category
fnwlgt	Integer
Education	Name of education level
Education-year	Numerical education level
Marital-Status	Category
Occupation	Category
Relationship	Category
Race	Category
Sex	Boolean
Capital-gain	Integer
Capital-loss	Integer
Hours-per-week	Integer
Native_country	Category
Salary	Binary (for prediciton >=$50k)

The features names are self-explanitory with the exception of *fnwlgt*. Maybe after working with the data I will be able to figure out what that is.

1.4 Data Cleaning and EDA

The data doesn't have a header, but the variables are listed on Kaggle. I'll add them here.

[ ]:
```
columns = [ 'Age', 'Workclass' , 'fnlwgt', 'Education', 'Education-year',␣
 ↪'Marital-Status',
            'Occupation', 'Relationship', 'Race', 'Sex' , 'Capital-gain',
            'Capital-loss', 'Hours-per-week', 'Native_country', 'Salary']

df.columns = columns
df.head()
```

[ ]:
```
   Age          Workclass  fnlwgt   Education  Education-year  \
0   50   Self-emp-not-inc   83311   Bachelors              13
```

```
1   38            Private  215646     HS-grad              9
2   53            Private  234721        11th              7
3   28            Private  338409   Bachelors             13
4   37            Private  284582     Masters             14

       Marital-Status          Occupation    Relationship    Race      Sex  \
0   Married-civ-spouse     Exec-managerial         Husband   White     Male
1             Divorced   Handlers-cleaners   Not-in-family   White     Male
2   Married-civ-spouse   Handlers-cleaners         Husband   Black     Male
3   Married-civ-spouse      Prof-specialty            Wife   Black   Female
4   Married-civ-spouse     Exec-managerial            Wife   White   Female

   Capital-gain  Capital-loss  Hours-per-week  Native_country  Salary
0             0             0              13   United-States   <=50K
1             0             0              40   United-States   <=50K
2             0             0              40   United-States   <=50K
3             0             0              40            Cuba   <=50K
4             0             0              40   United-States   <=50K
```

There don't appear to be any missing values to deal with.

[ ]:
```
df.isna().sum()
```

[ ]:
```
Age              0
Workclass        0
fnlwgt           0
Education        0
Education-year   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
Salary           0
dtype: int64
```

The *Education* and *Education-year* features are redundant. The year seems like it will be easier to work with, so I will drop the other feature.

[ ]:
```
df = df.drop(['Education'], axis=1)
df.rename(columns={'Education-year': 'Education'}, inplace=True)
df.head()
```

[ ]:

	Age	Workclass	fnlwgt	Education	Marital-Status	\
0	50	Self-emp-not-inc	83311	13	Married-civ-spouse	
1	38	Private	215646	9	Divorced	
2	53	Private	234721	7	Married-civ-spouse	
3	28	Private	338409	13	Married-civ-spouse	
4	37	Private	284582	14	Married-civ-spouse	

	Occupation	Relationship	Race	Sex	Capital-gain	\
0	Exec-managerial	Husband	White	Male	0	
1	Handlers-cleaners	Not-in-family	White	Male	0	
2	Handlers-cleaners	Husband	Black	Male	0	
3	Prof-specialty	Wife	Black	Female	0	
4	Exec-managerial	Wife	White	Female	0	

	Capital-loss	Hours-per-week	Native_country	Salary
0	0	13	United-States	<=50K
1	0	40	United-States	<=50K
2	0	40	United-States	<=50K
3	0	40	Cuba	<=50K
4	0	40	United-States	<=50K

For the next step I'd like to visuallize each feature against what I'll be trying to classify, Salary >= $50k. I'm most interested in Age, Education, Race, and Sex.