Finding Wealthy Customers - Case Study

Technical Specifications:

In this project I used;

- Numpy, Pandas for getting, cleaning and manipulating data.
- Matplotlib for plotting and visualization of data.
- Scikit-learn for the model. I used logistic regression because it is simple to use and based on my research it can usually be used for classification problems. In this problem, we are trying to find out that whether the income of the customer is more than 50K or not for people whose salaries are unknown. This is a classification, so we can use it here.

Process Details:

1) First of all, I read the data from csv file based on semicolons. Based on the data on the table there is total number of 45222 data, 11208 of them are having income more than 50K, and 28 455 of them are less than or equal to 50K.
2) Since we will get output from data, not NA or Null values, I dropped NA values from the data.
3) To see which columns have more unique values, I got value counts of all columns. Then I dropped all columns which more likely do not affect the income values. These are: 'hours-per-week' , 'capital-loss', 'capital-gain','age','native-country', 'Name'.
4) I mapped the values of the income column with 1 and 0 values.
5) Then mapped all the remaining string values with integer values. (you can see data_mapping.xlsx file.)
6) I grouped every column one by one and took the mean of income. We can see the independent attributes against income. We can see the outputs from the graphics below.
7) For the model, I used Logistic Regression from scikit-learn. I converted income to dataframe, and the other attributes to another dataframe by concatenating them. After that, I split them into random train and test subsets and performed train_test_split(df_income_x, df_income_y , test_size=0.33, random_state=42). (These values are based on default values/documentations).
8) Trained the model by training data, then I tried values with predict function.
9) To see the accuracy, I got **Accuracy: 0.780498318820588** by running;

```
print("Accuracy:",metrics.accuracy_score(y_test, y_predict))
```

!Note: After reviewing the attributes and accuracy rate, I decided to drop Name column as well, since it is affecting the prediction incorrectly.

Results:

Example 1:

Input → regression.predict([[1,7,3,6,0,2,0,11]]) :

Output → array([1])

Meaning: People who have the below attributes have more likely more than 50K income.

relationship	Wife	1
education_level	Prof-school	7
race	White	3
occupation	Exec-managerial	6
sex	Male	0
marital-status	Married-civ-spouse	2
workclass	Self-emp-inc	0
education-num	11	11

Example 2:

Input → regression.predict([[1,7,3,7,0,2,0,1]])

Output → array([0])

Meaning: People who have the below features have more likely less than 50K income.

relationship	Wife	1
education_level	Prof-school	7
race	White	3
occupation	Priv-house-serv	7
sex	Male	0
marital-status	Married-civ-spouse	2
workclass	Self-emp-inc	0
education-num	1	1

Example 3:

Input → regression.predict([[1,7,3,6,1,2,0,3]])

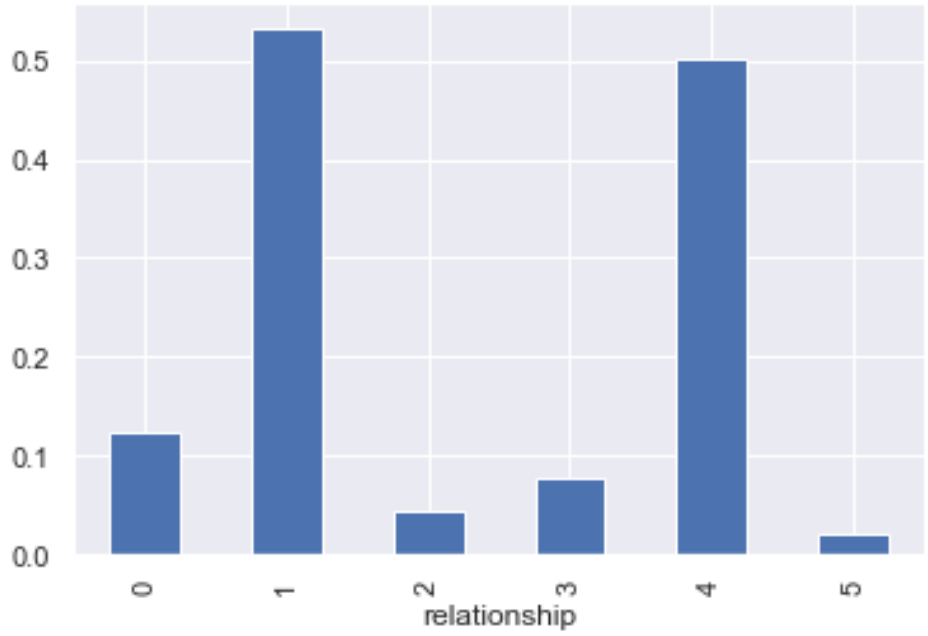
Output → array([0])

Meaning: People who have the below features have more likely less than 50K income.

relationship	Wife	1
education_level	Prof-school	7
race	White	3
occupation	Exec-managerial	6
sex	Female	1

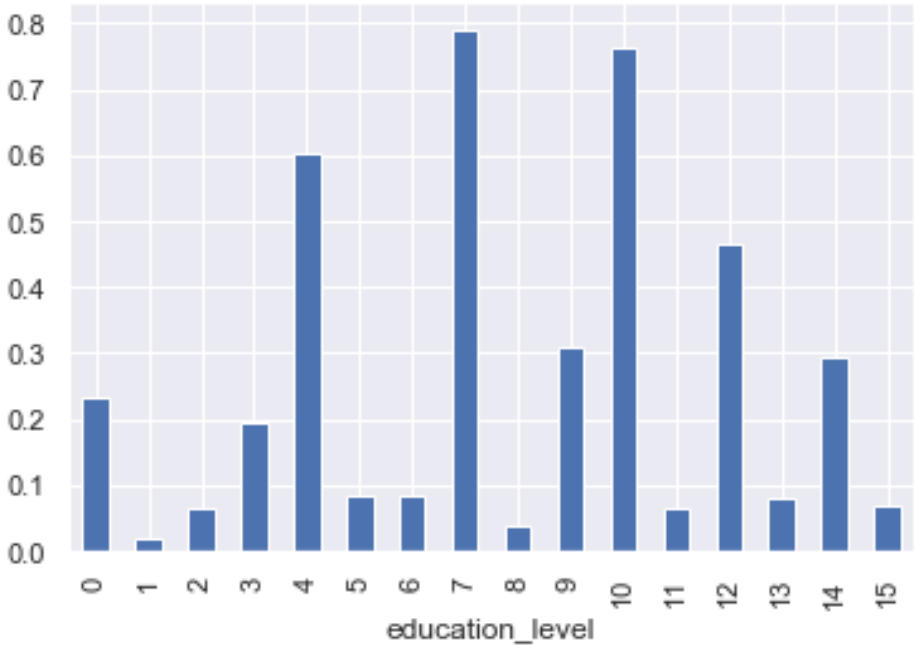
marital-status	Married-civ-spouse	2
workclass	Self-emp-inc	0
education-num	3	3

Outputs based on graphics (grouping by every attribute and taking mean of income):



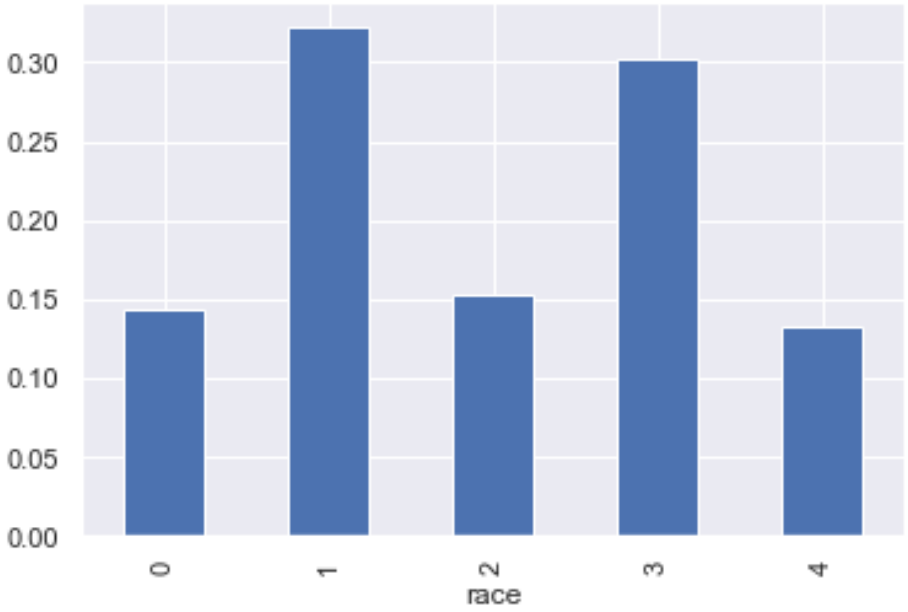
relationship	ID
Not-in-family	0
Wife	1
Other-relative	2
Unmarried	3
Husband	4
Own-child	5

- People that have Wife(1) relationship have income most probably more than 50K.
- People that have Husband(4) relationship have income most probably 50K.
- Others probably have less than 50K income.



education_level	ID
Some-college	0
Preschool	1
5th-6th	2
HS-grad	3
Masters	4
12th	5
7th-8th	6
Prof-school	7
1st-4th	8
Assoc-acdm	9
Doctorate	10
11th	11
Bachelors	12
10th	13
Assoc-voc	14
9th	15

- As you can see, people that have Masters (4), Prof-School (7) and Doctorate (10) education level most probably have income more than 50K.
- Others have probably less thank 50K.

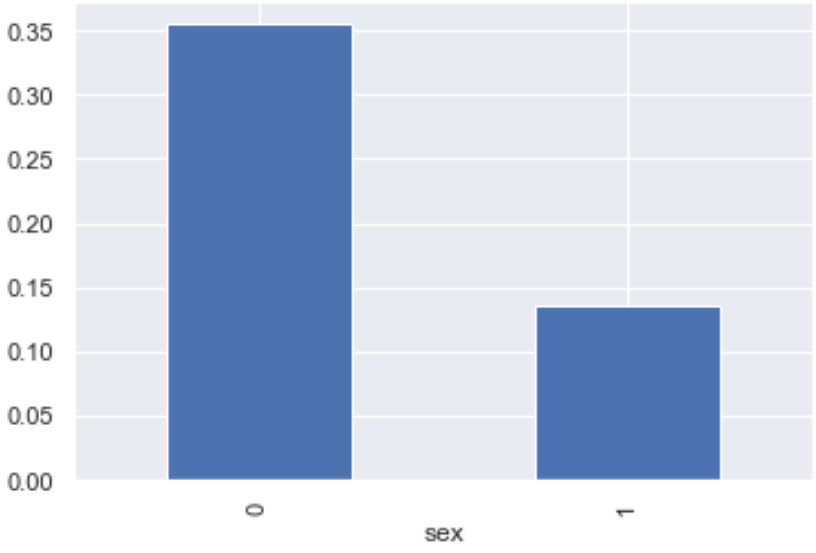


race	ID
Black	0
Asian-Pac-Islander	1
Other	2
White	3
Amer-Indian-Eskimo	4

- People whose race are Asian-Pac-Islander (1) and White(3) have more chance to earn more than 50K.

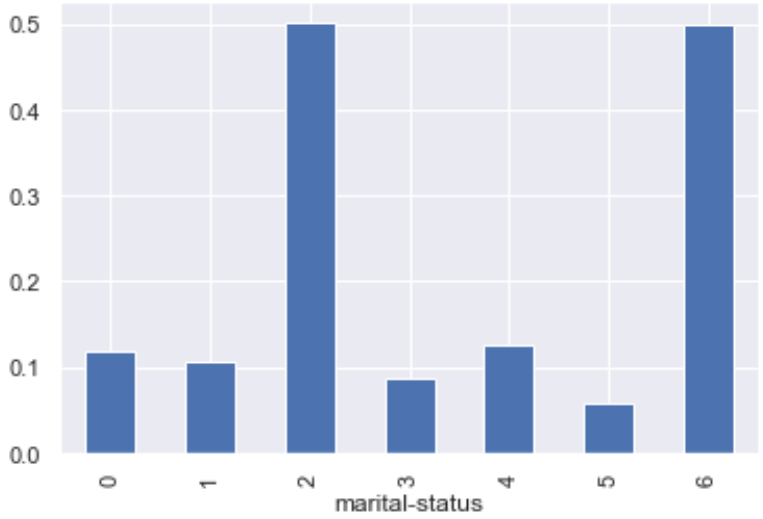
occupation	ID
Farming-fishing	0
Tech-support	1
Adm-clerical	2
Handlers-cleaners	3
Prof-specialty	4
Machine-op-inspct	5
Exec-managerial	6
Priv-house-serv	7
Craft-repair	8
Sales	9
Transport-moving	10
Armed-Forces	11
Other-service	12
Protective-serv	13

- People whose occupation is exec-managerial(6) have more probably more than 50K income.
- People whose occupation is prof- specialty(4) have more likely 50K income.



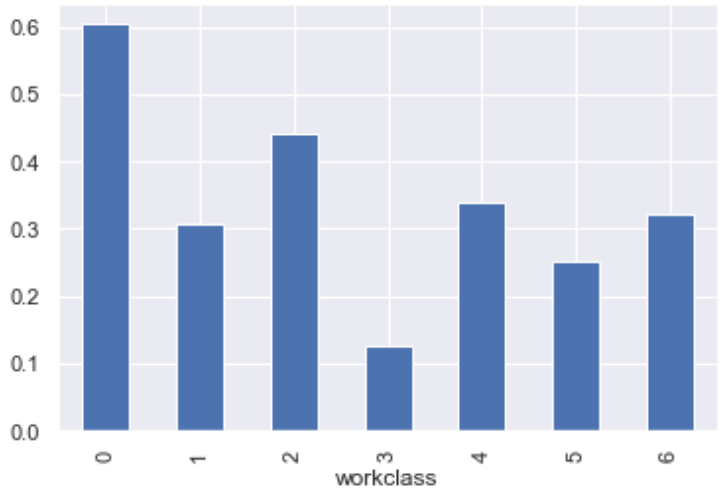
sex	ID
Male	0
Female	1

- As you can see people whose gender is male have more chance to have 50K income.



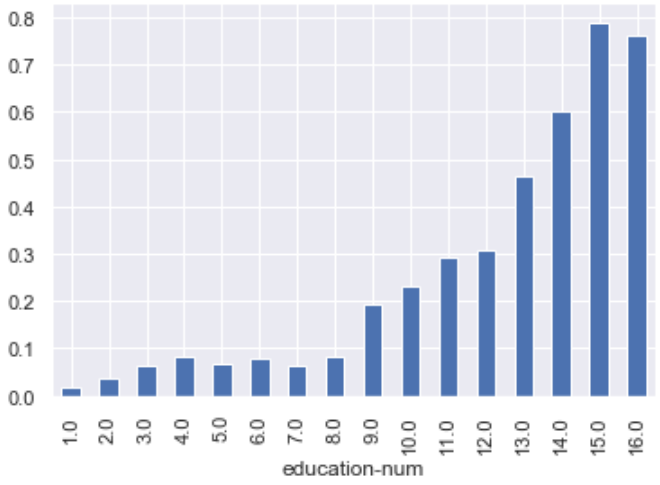
marital-status	ID
Married-spouse-absent	0
Widowed	1
Married-civ-spouse	2
Separated	3
Divorced	4
Never-married	5
Married-AF-spouse	6

- People who have the martial status “married-civ-spouse” and “married-AF-spouse” have more likely 50K income.



workclass	ID
Self-emp-inc	0
State-gov	1
Federal-gov	2
Without-pay	3
Local-gov	4
Private	5
Self-emp-not-inc	6

- People who have workclass “Self-emp-inc”(0) more likely to have more than 50K income.



- People who have education-num 14, 15, 16 have most likely more than 50K.

References:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://towardsdatascience.com/how-is-logistic-regression-used-as-a-classification-algorithm-51eaf0d01a78
https://towardsdatascience.com/logistic-regression-in-classification-model-using-python-machine-learning-dc9573e971d0

https://www.marktechpost.com/2019/06/12/logistic-regression-with-a-real-world-example-in-python/