COIMBATORE INSTITUTE OF TECHNOLOGY.

COIMBATORE.

Submitted by:

Nareshkumar G

1833029

Problem Statement 1:

1. Import the necessary libraries like pandas, Numpy, sns etc.
2. Reading the data with the help of pandas
3. Removing the column names of dataset so that it is easy to use
4. Separating the dataset into numeric dataset and non-numeric dataset
5. Summarising data
6. Doing EDA for our dataset, we found out the followings:

   Interpretation:

   - There are 6 numeric and 9 categorical columns.
   - There are 32561 rows and 15 columns.
   - Columns in the dataset are ['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occupation', 'relationship', 'race', 'sex','capital_gain', 'capital_loss', 'hours_per_week', 'native_country', 'target'] respectively.
   - We can see that 75% of values in capital_gain and capital_loss is zero, so we removed it.
   - Dataframe has 15 columns. There are 3 columns that have missing values. we can see that Occupation, workclass and native_country columns have few missing values
   - There are 24 duplicate Rows in the dataset. We should consider removing these rows.
   - It is clear that there is no strong correlation between the variables.
   - We can see that almost 75% of people might get below 50K of salary.
   - We can see that lot of people's occupation are Prof-specialty, Exec-managerial, Craft-repair and Adm-clerical and those who having Exec-managerial and Prof-specialty as their occupation are more likely to get an income of above 50k.
   - We can see that people are atleast completed their HS-grad and some college and those who completed their Doctorate and studied in professional school are more likely to get an income of above 50k.
   - We can see that there are lot more private workers than other category of workers and incorporated self-employees are likely to get salary of above 50k.
   - Average age for a person to get an income of above 50k is 39.
   - Elder people and males are more likely to get an income of above 50k.Clearly there are lot of noise data (Outliers) in age label.
   - Clearly, we can see that there are lot of older white peoples than other group of peoples.
7. Feature Engineering:
   - Replacing missing values with mode.
   - Checking if missing values are present then processed further.
   - Removing duplicates values.
   - Converting Categorical column to numeric column.
8. Feature Selection and spilt:
   - Splitting the training and testing set from both train and test datasets.
9. Feature Scaling:

- Scaling the train and test data so that every column is in certain range rather than influencing each other.
- By using RobustScaler(), we can remove the outliers as well as scaling the data.

10. Model fitting:
    - Here we use Logistic Regression because dependent variable(target) is discrete.
    - Model:
        - Output => 0 or 1: (<=50k or >50k)
        - Hypothesis => Z = WX + B
        - Activation function => Sigmoid (0,1)
        - Decision boundary => threshold = 0.5 (1 if y >0.5, 0 if y <0.5)
        - Cost function => Mean squared error (-y * log(h(x) - (1 - y) * log (1 - h(x))
        - Gradient descent => w = w – (learning rate* dw*T) & b = b – (learning rate* db)
          W – weight, b – bias.
    - Gradient descent updates the weights if cost function converges (minimize). There will be global minimum that mean where ever the point gradient descent starts it always converges at same point.
11. Predict the class 1 or 0.
12. Root mean Square error is calculated and RMSE is 0.44
13. Confusion matrix is build based on test set and predicted test set. Classification report is generated. Accuracy is around 80%.
14. Conclusion: Model we build classifies at decent rate.