**CENSUS INCOME DATA SET**

**INTRODUCTION**

The Adult Income Census data is a collection of 48,842 records obtained from the USA in 1994.

From our first segment, we analyze face value data to better understand relationships and correlations. Representation of some demographics in the corpus. This detail is then included in the segment Two frameworks to determine if a person made higher or less than $50,000 in 1994. Methods they use to gain insight into the same results. In the fourth segment, finally, Evaluate our templates as well as those of other within order to find out what the characteristicsare. Relevance, which approaches are most successful, and develop an understanding of some of them This is the feeling behind the figures.

**PROBLEM STATEMENT**

Adult Income Dataset, includes estimating private levels of income higher or lower than $50,000 per year on the basis of personal information such as marriage and occupation. There are far more cases of earnings below $50K than above $50K, but the bias is not serious.

This implies that unbalanced classification algorithms can be used while developed model can indeed be recorded using the classification accuracy, as is the case with balanced classification tasks.

The Adult Income Dataset seems to be from the Bureau Of labor statistics and the objective is to determine if a particular person earns upwards of $50,000 per year on the basis of characteristics such as occupation, duty hours each week, etc.

The dataset provides 14 input variables that are a mixture of categorical, ordinal, and numerical data types. The complete list of variables is as follows:

1.Age.

2.Workclass.

3.Final Weight.

4.Education.

5.Education Number of Years.

6.Marital-status.

7.Occupation.

8.Relationship.

9.Race.

10.Sex.

11.Capital-gain.

12.Capital-loss.

13.Hours-per-week.

14.Native-country.

The database includes incomplete values which are labelled with a question mark symbol (?). There seem to be binary classification values '>50K' and '<=50K,' i.e. a binary classification task. Groups are unbalanced, with a bias towards the mark '<=50K' class.

'>50K': dominant class, roughly 25%.

'<=50K': marginalized class, roughly 75%.

Utilizing predetermined training and testing sets, the estimated good classification error is about 14 per cent or the classification accuracy is roughly 84 per cent. This will provide a target to be set while developing on this data.

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

**DATA PRE-PROCESSING PIPELINE**

Following steps have been used to develop a machine learning model for this dataset :

1.Data Cleaning

2.Data Visualization

3.Data Pre-processing

4.Feature Engineering

5.Choosing the model

6.Fitting the model

7.Conclusion and results

**DATA CLEANING**

Again for categorical data, the null values are filled with the term "other." In the case of quantitative variables, we substitute the variables with the median, as the mean would offer a very skewed view. We end up with 1 dependent variable "income" and 12 estimating factors: "age," "job class," "education," "marital status," "occupation," "relationship," "race," "sex," "capital-gain capital-loss," "hour-per-week," and "native-country."

**EXPLORATORY DATA ANALYSIS**

The first plot itself shows the skewness of the output with majority of adults falling in the '<50k' category . This can be understdood easly using a countplot throught the seaborn function of python.

The second countplot plots gender wise plotting keeping the target as a basis .'hue' function is used here to plot gender wise segmentation of adults who fall into the two binary classifications of income . It can be observed that the skewness or the disparities in income occur more in the case of females compared to males

The third countplot shoes us the classification of people who fall into the binary classes based on their race and it can be observed that while the race ='white' shows high numbers in both the binary columns, however the count of other races in the '>50k' column is less.

The fourth countplot shoes us the classification of people who fall into the binary classes based on their relationship status and it can be observed that the '<50k' column is topped by adults with husbands and the '>50k' column is topped by adults who are not in families .

Histogram is a good technique to visually analyse the distribution of data and comparing disparities and skewness. So a histogram chart is used for all the numeric columns.

Next,a box plot is used on the data to view the outliers

**FEATURE ENGINEERING**

This step starts by replacing the question marks in columns like workclass, occupation and native-country with the most common value.

It can also be noticed that the education column has a lot of values which in a real life scenario might often be clubbed together so we have applied some feature engineering techniques and replace function to club multiple values under a broader head such as Preschool, Highschool and Higher ed. The code can be found below:-

df['education']=df['education'].replace(['Preschool','1st-4th','5th-6th','9th','12th','7th-8th','10th','11th',],'school')

df['education']=df['education'].replace('HS-grad','Highschool')

df['education']=df['education'].replace(['Assoc-voc','Assoc-acdm','Prof-school','Some-college'],'higher ed')

A similar feature engineering technique has been used for the column 'Marital Status' where similar values are clubbed under a broader relevant defination such as 'Married','Unmarried' and 'Others'. The code for this can be found below

df['marital-status']=df['marital-status'].replace(['Married-civ-spouse','Married-AF-spouse'],'Married')

df['marital-status']=df['marital-status'].replace(['Separated','Divorced','Widowed','Married-spouse-absent'],'others')

df['marital-status']=df['marital-status'].replace('Never-married','Not Married')

The next step is to actually quantify the target variable to numerical format so that we can evaluate and quantify our binary classification.

Also, a function called the standard scaler is used to scale the data and bring it to a similar evaluation metrics so that the data is in one standardized format. This plays a huge role in accuracy of the model

**THE MODEL**

The aim of this phase is to create a model that supports us as the basis for measuring the output of a stronger and more tailored model. We use various classification Techniques and compare them to see which model provides better results

The whole data is divided into X and Y where X has all the variables except the target which is 'Income' and Y only has one feature which is 'Income'.

I have used three models here namely:

1.Logistic Regression-81 percent accuracy

2. Support Vector Classifier-85 percent accuracy

3.Decision Tree Classifier-81 percent accuracy

Based on the accuracy of the models , we have come to a conclusion that SVC is the best model for this data among the three.

This step is followed by drawwing AUC-ROC curves and printing the comfusion matrix

**CONCLUDING REMARKS**

Feature Selection and engineering is the most crucial thought in this type of issue. You will see how we've treated numeric and categorical data and how we've developed a different Machine learning framework on the very same sample group. We also review the accuracy score score of each method so that we can explain how it should be done in our test dataset. Finally, you could also help develop the Model by tuning the multiple metrics used in the model.

Logistics Regression has been the least effective because the data had a major weakness causing inconsistency. The manner the classifier functions. The binary data points suggest inadequate preparation in the Second Y vector and also has problems of not being absolutely linearly separable.

In this blog, we discussed a variety of models and how to apply them. Classification model dataset, resolving a variety of concerns such as feature engineering, data disparity followed Model building and hyper-parameter adjustment to improve overall the accuracy of the results. I can evidently see information being skewed, but our predictive resources have been limited. When we wanted to oversample the minority, we managed to do so normally.