MACHINE LEARNING FOR DATA ANALYTICS

STEPHEN ALGER - DT228/4
C16377163
OCTOBER 2019

Introduction to the Dataset - What does this data represent?
The dataset provided is 1994 Census data from the US, with 30940 entries. It is being analysed in order to form models by which we can predict whether or not an individual made ±50,000$ in a year.

This Dataset consists of the following data points (features) about each individual:
['id', 'age', 'workclass', 'fnlwgt', 'education', 'education-num', 'marital-status', 'occupation', 'relationship', 'race', 'sex', 'capital-gain', 'capital-loss', 'hours-per-week', 'native-country', 'target'].

In my code, I separate categorical and continuous features as required. Then we can compile our respective ABT Reports for both Data Types: Categorical & Continuous. We can also then graph our data as I have done in my code. Continuous Data suiting histogram graphing and categorical suiting bar charts.

Cardinality - I believe Irregular cardinality is not a plaguing issue with this dataset in my report, with the only real feature to consider being 'fnlwgt' - 'final weight' which has 20880 distinct values in this dataset as shown in my ABT Report. This feature represents the number of people believed to be represented by a single entry. From my review of the Data Quality report the rest of the features have a regular cardinality consistent with possible values in their respective census questions.

Missing Values - While our ABT (Analytical base table) immediately confirms the continuous set of data was found to contain zero missing values, the same cannot be said for the Categorical aspect of the data. Our ABT highlights the features and their missing percentages - which should give cause for concern are the following: {workclass: 5.61%, occupation: 5.63%, native-country: 1.79%}.

Missing Values are a major DQI (data quality issue) which affect the accuracy of the rest of the ABT produced and thus weakens the value of any derived models from the dataset. For example in our Dataset, if the 5.63% of missing values of the Occupation feature were collected as intended - it could easily have a material affect on our 1st & 2nd Modes and invalidate them, due to the small modal frequency (45) differential between them making Craft-repair the modal value of the Occupation feature. In this dataset missing values were replaced with a 'NaN' identifier however techniques such as imputation with a measure of central tendency.

Outliers - Reviewing the Continuous ABT report it is obvious that there are several instances of outlying values - particularly in the Capital Gain and Loss features where the maximum is perhaps inconsistent with the statistical metrics. In both cases the Min, 1st Quartile, Median (2nd Quartile), 3rd Quartile are zero yet the maximum is significantly higher. From this we can see a large range in the data, from this the resulting Standard Deviation is very high -> this concerns me as it greatly reduces our ability to predict accurate outcomes. note: (Capital Gain Range: 0,99999).

From the ABT Categorical Report we can judge that we have a few very dominant modal outcomes, showing this strong response of particularly White, Male, US Born respondents. Other obvious points we can infer from the data are that the central tendency statistics of the Continuous ABT report show the overwhelming majority of people self report to work 40 hours per week. I would definitely question the gathering techniques of this census also with values such as Hours Worked per week, capital gains etc being presumably self reported and unverified allowing for high degrees of bias and inaccuracy.

Interesting Points which will be very useful to form a ML model:
- Age Distribution with respect to Income
- Education Distribution) with respect to Income
- Working Class distribution with respect to Income
- Hours Worked Per Week VS Income
- How does Sex affect Earning
- How does Race affect Earning
- How does Earning differ among Occupations