Loughborough University

COP259: Data Mining Coursework

Credit value: 100% of the module

This coursework calls for a report on the tasks that you are to carry out giving results and their interpretation on a data set that you have been given to analyze using the Orange Data Mining software.

Please note, this is a strictly individual assignment. You are not allowed to collaborate on any aspect of the work.

In this assignment you are not expected to write any software code but use existing tools provided by Orange.

Perform the following tasks on the dataset provided and produce results and their interpretation in the form of a report.

Please note:

- You must only use the data set(s) provided.
- In-app screenshots are compulsory to show the result of each part (not step by step illustrations).
- Word count for each part should not exceed 500 words.

Your task

In this assignment you will carry out some exploratory analysis of census data from the United States. This type of dataset is an anonymized subset of a census study that has been made available for public use. Such datasets in both the UK and US are known as a Public Use Microdata Sample (PUMS) and they are very useful for educational and basic research purposes. The file you will work with is ***Census_data.csv*** which you can find on Learn, and it contains the following tabulated attributes:

- Age
- CoW (category of work)
- Marital (status)
- Occupation
- PoB (place of birth)
- Hours (weekly hours of work)
- Sex (male or female)
- Race
- State
- Income

Part 1. Preprocessing

☐ As the dataset is quite large, to speed up computations, reduce to 5000 data points.

In the current state of that file attributes are numerically coded to save space. E.g. the state of Alabama is coded as 1, Sex: Female is coded with 2 etc. Before we can perform any meaningful processing, we must transform the data to a readable form. To that end, you are given file ***Attribute_values.csv*** (also to be found on Learn) which consists of triplets of the form

- Attribute name (the name of the attribute)
- Attribute numeric value (a numeric code)
- Attribute value (the actual value corresponding to that code)

☐ You should end up with the following data structure:
  - CoW: Private Employee, Government Employee, Self-Employed, No pay, Unemployed.
  - Education: keep the numeric value (it corresponds roughly to years in full time education) but create a new attribute with the text descriptions, making sure to group categories 1-15 as “no diploma”, 16-17 as “high-school”, 18-19 as “post-high-school” and keeping the rest as they are in Attribute_values.csv
  - Marital: one of Single, Married, Widowed, Divorced, Separated
  - Occupation: keep the number but create a new attribute with the type of industry (Hint: this can be conveniently found in the first three characters of the long description of occupation). May come handy for task 6.
  - Place of birth: we are only interested in US/non US so please group into two categories accordingly.
  - Sex: replace with “male”, “female”
  - Race: turn this into a binary variable consisting of “white” and “non-white” labels
  - State: replace with name of state.

☐ Handle any missing data or outliers

☐ Perform exploratory data analysis to understand the dataset’s structure and key characteristics. Include visualizations and summary statistics.

[20 marks]

Part 2. Fairness in income distribution

Here you will do some analysis on the distribution of income in the US.

- ☐ Plot a histogram of income.
- ☐ Plot a histogram of log-income.
- ☐ Create a zipf plot for income. This is a plot where the Y axis is log-income and the X axis is the log of the rank of each income in the whole distribution. E.g. the highest income has rank 1, the second highest has rank 2 etc. *[Hint: for this question you might find it easier to create a ranked list of incomes using a different software e.g. Excel. Alternatively you can use the create timeseries using negative income as the time variable.]*
- ☐ How would you interpret these results?

You will investigate whether sex, race and place of birth affect one's income. To simplify the analysis, for race you can consider a simplified category that only considers White vs Other.

- ☐ Visualize the income distribution broken down by each of these attributes.
- ☐ Compute a t-test to verify the statistical significance of any biases you find.
- ☐ Comment on your findings.

Investigate correlations between income and (1) age, (2) hours worked/week and (3) education level (use the numeric value here).

- ☐ Create scatter plots of the pairs above
- ☐ Compute Pearson correlations
- ☐ Repeat the two steps above for log-income
- ☐ Comment on your findings and on any differences between income and log-income

[20 marks]

Part 3. Predicting income

In this part you will use various models to identify the key factors that determine one's income. Firstly, you will explore the relationship between income and education:

- ☐ Plot *mean income* against *education level*
- ☐ Estimate a monetary value for education, by estimating how much one's income increases per year of education.
- ☐ Describe what could go wrong with the analysis above.

You will now apply some classification models to predict incomes.

- ☐ Simplify things by splitting the population in the middle according to income. Let's call these sub-groups "low income" and "high income" respectively.
- ☐ Try out several different classifier models (at least 5) to predict which subgroup a person belongs to.

Use the best model you found for the remainder of this analysis.

- ☐ Do feature ranking using the income subgroup as a target. Make a note of the results.
- ☐ Train the model on your data and apply at least two different model explainability techniques (e.g. using SHAP and permutation-based, feature importance)
- ☐ Do the results agree with the feature rankings obtained previously?
- ☐ Comment on your findings

[20 marks]

Part 4. Demographics of US elections

For this part you will use ***voting_2020.csv*** which contains election results by state for the 2020 (Trump vs Biden) election.

- ☐ Plot the election results, mean income and mean educational attainment level on the US map.
- ☐ Comment on the visual comparison of the maps
- ☐ Use any visualization or statistical test to examine the hypotheses that
    - "low income states voted for Trump" and
    - "states with high educational attainment voted for Biden."

[20 marks]

Part 5. Your own data mining

- ☐ Formulate a hypothesis on your own, that is both interesting and non-trivial and examine its validity using the datasets provided. Use any relevant data mining techniques and apply rigorous statistical testing where possible. Justify any choices you make. Examples of hypotheses include:
    - "Higher education levels lead to higher income, but the effect varies significantly across different states."
    - "The impact of weekly working hours on income is moderated by the type of occupation."
    - "There is a significant difference in income distribution between urban and rural areas."

[20 marks]