Machine Learning and Optimization - CT7205
Assignment

This assignment has two questions. Each of questions has a set of exercises starting from data exploration to model building and evaluations. The main purpose is to test your abilities, as a data analyst, in solving real-life machine learning problems by following a proper methodology.

The deadline for this assignment is Friday 22rd July at 3pm.

You are given all the data in csv files. You need to submit ONE document only (word file/PDF file/ Jupiter notebook). You need to add the code (screen shot is accepted) and discuss the code in your report. It is recommended to add comments in your code.

Question 1. Medical Insurance (40 % weight)

Nowadays, medical insurance is becoming a necessity for many people. Depending on the medical care, insurance companies collect annual premiums. It is difficult to estimate the medical expenses due to various health conditions of payees. Some conditions are, however, more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

As a result, insurers invest a great deal of time and money to develop models that accurately forecast medical expenses. As a data scientist, you were given some real-life patient data in (insurance.csv) with the following 7 columns:

- age: age of primary beneficiary
- sex: insurance contractor gender: female or male
- bmi: body mass index, providing an understanding of body, weights that are relatively high or low relative to height
- children: number of children covered by health insurance
- smoker: yes or no
- region: the payees' residential area in the US, northeast, southeast, southwest, northwest
- medicalCost: individual medical costs billed by medical insurance.

Your task is to use machine learning to build a prediction model to estimate the medical cost of individuals based on the predicters given in the dataset. Use the data science methodology to achieve the task. The following questions need to be answered

a) Is the required ML supervised, unsupervised, or semi supervised learning and why? Which ML task (classification, clustering, regression analysis or any other) is the best in this case and why?
b) Explore your data and document your observation.
c) Study the correlation between each predictor and the medicalCost. What is your conclusion?
d) Use the correlation analysis to select 3 best predictors and build a simple linear regression model based on each of the predictors.
e) Evaluate the performance with the statistical performance measures to evaluate the statistical significance of your results.
f) Build two multivariate regression models 1) with the three predictors above and 2) with all the predictors in the dataset. Evaluate and compare the two models.
g) State your overall conclusions for this task.

Question 2. Census Income (60 % weight)

In this exercise, we use the US Census dataset from the Census bureau (publicly available from UCI Machine Learning Repository). The task is to predict whether a given adult makes more than $50,000 a year or not based on a number of attributes.

The dataset has 14 column names as below:

- age: the age of an individual
- workclass: employment status of an individual
- fnlwgt: final weight. In other words, this is the number of people the census believes the entry represents
- education: the highest level of education achieved by an individual
- education-num: the highest level of education achieved in numerical form
- marital-status: marital status of an individual.
- occupation: the general type of occupation of an individual
- relationship: represents what this individual is relative to others •
- sex: the biological sex of the individual
- capital-gain: capital gains for an individual
- capital-loss: capital loss for an individual
- hours-per-week: the hours an individual has reported to work per week •
native-country: country of origin for an individual
- label: whether or not an individual makes more than $50,000 annually.

Your task is to use machine learning to build a machine learning (ML) models. You can apply some assumptions if necessary, but you need to explain what assumptions and why you applied in your ML model? Again, use the data science methodology to achieve the task. The following questions need to be answered.

a) Load and explore the data (note your observations).

b) Use appropriate methods to handle categorical data

c) Investigate and train at least 5 ML models including Classification (to predict if an individual going to earn more $50,000 annually or not), Clustering and Neural Networks. You are free to choose any ML algorithms.

d) Optimise your models, evaluate the models and compare the models' results as:

   i. How optimisation improve the performance of the model? Which parameter you used for optimisation.

   ii. Compare the results among the models of similar types (eg. If you using two classification models, compare their performances)

e) State your overall conclusions for this task.

Please feel free to contact me if you have any questions bmishra@glos.ac.uk

****** Good Luck ******