BC2407 Specimen Paper for CBA Revision

Total Marks:	100
Type:	Do-at-home individual assignment

This paper contains 6 questions across 5 pages, including this cover page.

US Health Insurance Charges

Introduction

"Health care in the United States can be very expensive. A single doctor's office visit may cost several hundred dollars and an average three-day hospital stay can run tens of thousands of dollars (or even more) depending on the type of care provided."

--- https://vaden.stanford.edu/insurance/health-insurance-overview/how-us-health-insurance-works

Thus, health insurance is a common way for individuals to pay for health care. The insurance premiums (up-front costs) and co-payments (upon utilization of health care) could vary significantly depending on the individual risk and type of insurance policy.

A set of data is sourced from Kaggle[1] to understand how insurance charges vary with different individual profile. The data dictionary is also provided in Appendix A.

Answer the following questions in word document with your class and name as the filename. Check that only three variables listed below are categorical throughout the analysis:

- sex
- smoker
- region

Questions

Part A: Data Exploration (20 marks)

1. Explore the relationship (if any) of each variable to charges. Which variable(s) seems to be important in explaining charges? (in less than 100 words.)

[1] Source: https://www.kaggle.com/teertha/ushealthinsurancedataset

Part B: Predictive Techniques (30 marks)

2. Do a 70-30 train-test split and verify with screenshots that all the categorical levels in the 3 categorical variables are represented in both the trainset and testset.
3. Develop 3 models to predict charges:
    a. Backward Elimination Linear Regression
    b. MARS
    c. Random Forest

   Report the testset RMSE (round off to nearest dollar) in the form of a table:

Model	Testset RMSE
Backward Elimination Linear Regression	
MARS	
Random Forest	

Part C: Insights and Opinion (30 marks)

4. Based on the 3 models developed in Part B, explain your insights in less than 200 words. Provide screenshots of software output to justify your insights and/or opinion.

Part D: Limitations and Enhancements (20 marks)

5. Explain the limitations of this analysis (in less than 100 words).
6. Suggest ways to enhance the analysis and insights (in less than 100 words).

Appendix A

Data Dictionary

age:
Age of primary beneficiary.[2]

sex:
Insurance contractor gender, female / male.

bmi:
Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9.

children:
Number of children (aka Number of dependents).

smoker:
Smoker / Non – smoker.

region:
The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

charges:
Individual medical costs billed by health insurance.

Source: https://www.kaggle.com/teertha/ushealthinsurancedataset

[2] Typically, the insured person is the primary beneficiary in medical or hospitalization insurance.