UNIVERSITY OF EXETER BUSINESS SCHOOL

Homework Assignment

Machine Learning for Economics

BEE3066

Maximum Marks: **100**

Deadline: *November 19, 2024, 15:00*

Before the end of deadline you must submit a single PDF document containing your answers to the questions, including any tables and graphs. At the end of the document include the R code you used to obtain the answers in an appendix.

Materials to be supplied: Data file.

Detailed Instructions:

Answer all questions. Feel free to use all course material, textbook, etc. to complete this homework.

Submission is online i.e. you have to upload the PDF document on ELE website. The submitted document will contain your answers to the questions, including any output, tables and graphs. Also, include the R code containing the commands you used to obtain your answers in the same document at the end as an appendix. You must indicate which question your commands relate to.

You may only submit once; it is not possible to alter your document and re-submit at a later time.

The data corresponding to this homework is uploaded on the ELE (*insurance.csv*).

Marking will follow the university's assessment guidelines.

Introduction

A medical expenditure data set is provided (*insurance.csv*) at the individual level. It contains information on individuals (subjects) who are 65 years and older who qualify for the U.S. Medicare program. This program does not cover all expenses. For example, if some treatment (or drug) is not part of the U.S. Medicare program, it will not be paid for. Hence, some of the individuals buy a supplementary private insurance that provides insurance coverage against various out-of-pocket expenses. The data contains the following variables:

Variable	**Description**
sid	Subject ID
age	Age
famsze	Size of the family
educyr	Years of education
totexp	Total medical expenditure
retire	=1 if retired
female	=1 if female
white	=1 if white
hisp	=1 if Hispanic
marry	=1 if married
northe	=1 if North-East area
mwest	=1 if Mid-West area
south	=1 if South area (West is excluded)
phylim	=1 if has functional limitation
actlim	=1 if has activity limitation
msa	=1 if metropolitan statistical area
income	annual household income (in 1000 dollars)
injury	=1 if condition is caused by an accident/injury
priolist	=1 if has medical conditions that are on the priority
totchr	# of chronic problems
suppins	=1 if has supplementary private insurance
hvgg	=1 if health status is excellent, good or very good

Questions

1. [1 mark] How many observations are there in the data?

2. [3 marks] Which 5 variables are most correlated with the total medical expenditure (*totexp*)?

3. [5 marks] Plot five graphs each with one of the 5 variables identified in (2) on the x-axis and *log(totexp)* on the y-axis. Discuss each graph briefly. Make sure to change all indicator variables in the data to factor (categorical) variables.

4. [38 marks] Estimate a linear regression model with $log(totexp)$ as the dependent variable and $suppins, phylim, actlim, totchr, age, female, income, mwest$ and $northe$ as the independent variables. Based on this linear regression answer the following:

   (a) [2 marks] Is the model overall significant?

   (b) [4 marks] How much is the residual standard error? Interpret it.

   (c) [4 marks] Interpret the coefficients on the $female$ and $income$ indicator variables.

   (d) [8 marks] Produce diagnostic plots for the linear regression and briefly discuss each of the four graphs.

   (e) [4 marks] How many observations can be classified as outliers and high leverage points (three times the average leverage)?

   (f) [5 marks] How does the regression output change if we remove the outliers?

   (g) [5 marks] How does the regression output change if we remove the high leverage points?

   (h) [6 marks] Estimate the linear regression model after removing the variables which are not strongly significant ($p-value > 0.01$). Is the new model superior or inferior in terms of fitting the data?

5. [36 marks] Construct a new indicator variable, *high*, which equals 1 if the *totexp* of an individual is above its median value, 0 otherwise. Add the *high* variable to the data frame as a factor variable. Answer the following:

   (a) [10 marks] Split the data into training and test data: Randomly sample 2000 observations for the training data. Rest of the observations will form the test data. Perform logistic regression on the training data to predict *high* using $income, actlim, phylim$ and *totchr* as the independent variables. Calculate the test error rates for the following cut-offs of the estimated probability: 0.2, 0.4 and 0.6. Which is the most preferred cut-off among the three?

   (b) [6 marks] Calculate the test classification error rates for the LDA method using the following cut-offs of the estimated probability: 0.2, 0.4 and 0.6.

   (c) [6 marks] Calculate the test classification error rate for the QDA method using the following cut-offs of the estimated probability: 0.2, 0.4 and 0.6.

(d) [6 marks] Calculate the test classification error rate for the KNN classifier using the following values for K= 1, 5, 10, 20, 50, and 100. What is the most preferred value of K to minimise the overall error rate?

(e) [8 marks] Which method appears to provide the best results on this data? What can we infer about the data generation process?

6. [7 marks] Compute the standard errors for the LDA coefficients on *income*, *actlim*, *phylim* and *totchr* variables using the bootstrap method. Use the complete data and the same model specification as in the question (5 – (b)). Create 95% confidence interval around the LDA coefficients using the standard errors obtained from the bootstrap method. Assume, Confidence Interval = Coeff. ± 1.96 × Std. Error.

7. [10 marks] Calculate the LOOCV error and the k-fold CV error for the logistic regression model at a cutoff of 0.3 for $k = 2, 5, 10, 20$ and 50. Use the complete data and the same model specification as in the question (5 – (a)). Compare and discuss LOOCV and k-fold CV estimates of the test error?

——— *End of Assignment* ———