UNIVERSITY OF EXETER BUSINESS SCHOOL

Homework Assignment

Machine Learning for Economics

BEE3066

Maximum Marks: 100

Deadline: *November 16, 2023*

Submit a single PDF document containing your answers to the questions, including any tables and graphs, before the deadline. At the end of the same PDF document, please include the R/Python code you used to obtain the answers.

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/Word document on BART website. The submitted document will contain your answers to the questions, including any output, tables and graphs. Also, include the R or Python code containing the commands you used to obtain your answers in the same document at the end. 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 page (*labour.csv*).

Marking will follow the university's assessment guidelines.

Introduction

You are given a data set (*labour.csv*) to study the effect of labour union membership on wages. The data contains information on 545 individuals over 8 consecutive years. Hence, the total number of observations is 4,360. The indicator variable, *union*, has value 1 if the individual holds the labour union membership in a particular year, 0 otherwise. The data contains the following variables:

Variable	**Description**
nr	Unique ID of an individual
wage	Income earned (£1000)
year	Year
black	=1 if the individual is black
hisp	=1 if the individual is hispanic
exper	Work experience (#years)
hours	Hours worked in a year
married	=1 if the individual is married
occ1	Occupation Cateogry 1
occ2	Occupation Cateogry 2
occ3	Occupation Cateogry 3
occ4	Occupation Cateogry 4
occ5	Occupation Cateogry 5
occ6	Occupation Cateogry 6
occ7	Occupation Cateogry 7
occ8	Occupation Cateogry 8
occ9	Occupation Cateogry 9
educ	Number of school years
union	=1 if individual has labour union membership

Questions

1. [5 marks] Compute the correlation matrix for all variables in the data. Which 4 variables are most correlated with *wage*? Produce a pairwise scatterplot with *wage* and these 4 variables. Briefly interpret the scatterplots.

2. [5 marks] Change all indicator variables in the data ($black, hisp, married, union, occ1 - occ9$) into factor (categorical) variables. Graph a box plot with the variable *married* on the x-axis and *wage* on the y-axis. Discuss the graph briefly.

3. [8 marks] Estimate a linear regression model with *wage* as the dependent variable and all other variables as independent variables except $occ9$. Answer the following:

   (a) [2 marks] Is there a relationship between *wage* and all the predictors?

   (b) [3 marks] How much is the residual standard error? Interpret it.

   (c) [3 marks] Interpret the coefficients on the *married* and *hisp* indicator variables.

4. [25 marks] Re-estimate the linear regression model with *wage* as the dependent variable and including only those variables which were significantly associated (p-value < 0.1) in the last question. Based on this regression model, answer the following:

   (a) [8 marks] Produce diagnostic plots of the linear regression. Briefly comment on each of the four graphs.

   (b) [6 marks] How many observations can be classified as outliers and high leverage points (three times the average leverage)? What can we infer from the difference between the number of outliers and number of high leverage points?

   (c) [5 marks] Compute the variance inflation factor for all the predictor variables. What can we infer from it?

   (d) [6 marks] The coefficient on the *hours* variable is negative. Is it expected? Explain.

5. [37 marks] Construct a new indicator variable, *rich*, which equals 1 if the *wage* of an individual is above the mean *wage*, 0 otherwise. Add the *rich* variable to the labour data frame and change it to a factor variable. Answer the following:

   (a) [10 marks] Split the data into training and test data: test data will contain observations from year 1980, 1981 and 1982. Rest of the observations will form the training data. Perform logistic regression on the training data to predict *rich* using $educ, hours, exper, black$, and *married* as the independent variables. Calculate the test error rates for the following cut-offs of the estimated probability: 0.4, 0.6 and 0.8. 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: 0.4, 0.6 and 0.8.

(c) [6 marks] Calculate the test classification error rate for the QDA method using the following cut-offs of the estimated probability: 0.4, 0.6 and 0.8.

(d) [7 marks] Calculate the test classification error rate for the KNN classifier using the following values for K: 1, 10, 20, 50, 100 and 200. 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? Discuss.

6. [8 marks] Compute the standard errors for the LDA discriminant coefficients on $educ, hours, exper, black$, and $married$ variables using the bootstrap method. Use the complete data and same model specification as the last question (5 – (b)).

7. [12 marks] Calculate the LOOCV error and the k-fold CV error for the logistic regression model at a cutoff of 0.6 for k = 10. Use complete data and same model specification as the earlier question (5 – (a)). Compare and discuss LOOCV and k-fold CV estimates of the test error?

——— *End of Assignment* ———