*This is the main submission document****. Save and rename this document filename with your registered full name as Prefix before submission.***

|  |  |
| --- | --- |
| Class | 1 |
| Full Name | AARON FOO CHUAN TONG |
| Matriculation Number | U1821884B |

*\* : Delete and replace as appropriate.*

**Declaration of Academic Integrity**

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University’s Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

*Please insert an “X” within the square brackets below to indicate your selection.*

**[ X ] I have read and accept the above.**

Table of Contents

[Answer to Q1: 2](#_Toc87111201)

[Answer to Q2: 3](#_Toc87111202)

[Answer to Q3: 4](#_Toc87111203)

[Answer to Q4: 5](#_Toc87111204)

[Answer to Q5: 6](#_Toc87111205)

[Answer to Q6: 7](#_Toc87111206)

*For each question, please start your answer in a new page.*

# Answer to Q1:

Create the BMI variable based on CDC definition1. Show your code.

[Start your answers here…]

Creation of the BMI variable , BMI = weight in Kg / height in meters ^2

Text

Description automatically generated

Table

Description automatically generated

# Answer to Q2:

There are many categorical variables with integer coded values (e.g. Diabetes, HighBloodPressure, Transplant…etc.) Is it necessary to convert them to factor datatype in R?

[Type your answers here…]

It would not be necessary to convert these variables into factor datatype as they as are True/False variables represented as 1’s and 0’s in the csv file. Thus, there is no need to order them as it would not provide any additional information unlike the ordering of ordinal variables like shirt size or time of day which when ordered could allow models to learn from the patterns.

# Answer to Q3:

Explore the data and report on your key findings.

[Start your answers here…]

A screenshot of a computer

Description automatically generated with medium confidence

From the summary we note that the mean age is 41.72 years old, and the average person is overweight with a BMI of 27.45.

High blood pressure, diabetes and having a major surgery are relatively common with occurrence rate of > 40%.

Having a transplant is the rarest with occurrence rate of ~5%.

Table

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

From the correlation plot we can observe for the premiums row that mainly age, high blood pressure, transplant, chronic disease, number of major surgeries and weight/BMI have a positive correlation with insurance premiums.

The clearest relationships are age, transplant, and number of major surgeries with correlations of:

Premium and age: 0.70

Premium and transplant: 0.29

Premium and number of major surgeries: 0.27

With age being the clear leader in determining premium payments.

# Answer to Q4:

Using 1 SE optimal CART and one other technique learnt in this course:

a. What is the 10-fold cross validation RMSE and number of splits in the 1SE Optimal CART?

b. Identify the key predictors of premium.

c. Is BMI or Gender important in determining premium?

d. Evaluate and compare the predictive accuracy of the two techniques on a 70-30 train-test split. Present testset RMSE results in a table.

[Start your answers here…]

1. 10 fold cross validation RMSE is:

Table

Description automatically generated with medium confidence

Text

Description automatically generated

As a chart:

Chart

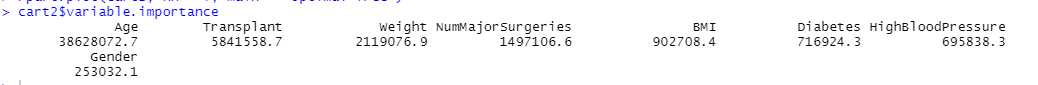
Description automatically generated

Number of splits for 1SE Optimal CART is: 5 splits

A picture containing text

Description automatically generated

b) Key predictors



From variable.importance we can tell that Age, Transplant and Weight are the top 3 predictor variables in affecting the model’s performance

c) BMI is shown to be more important than Gender in determining premium. Its importance factor is more than 3 times that of gender ( 902708/253032 =3.56).

d) **CART**

Graphical user interface, text

Description automatically generated

**LINEAR REGRESSION**

Graphical user interface, text, application

Description automatically generated with medium confidence

Summary of results:

|  |  |  |
| --- | --- | --- |
|  | CART | LINEAR REGRESSION |
| RMSE | 182 | 174 |
| MAPE | 4.86% | 11.7% |

# 

# Answer to Q5:

Explain the limitations of your analysis. [Max 1 page.]

[Start your answers here…]

RMSE being the sum of squared errors would penalise harder larger deviations from the actual values. Both the linear regression model and CART model were then set to train on minimising the sum of squared errors as a signal to train the model.

However, this number is an absolute value and does not give us a sense of proportionality of the errors in comparison with actual values unlike MAPE (mean absolute percentage error). Additionally, no consideration on the usability of the models was given. A pruned CART model can be very readable, fast and user friendly when printed as compared to inputting values into a Rscript in order to get the predictions from a linear regression model.

A picture containing diagram

Description automatically generated

As we can see here the pruned CART decision tree provides a simple visual chart which can be used by anybody without any knowledge of R code.

# Answer to Q6:

Is CART successful in this application? Explain. [Max 1 page.]

[Start your answers here…]

Yes. CART provides a simple solution to the complex problem of estimating insurance premiums for potential customers. By using CART the insurance company will be able to let visitors identify for themselves the risk group they belong too and determine their most probable insurance premiums when engaging with this insurer. This will help reduce the workload of the staff as customers are now ‘informed parties’ and already know what they want when seeking an agent of the company. Through CART we greatly reduce the probability that customers consult our agents and upon finding out the premium offered, walk away from the company.