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

Full Name	CHEN LUYU
Email Address	CHEN1819@e.ntu.edu.sg

** : 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

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

Answer to Q1:

1. The dataset reflects the LOS of patients who are all diagnosed as Septicemia, which means that the prediction result may not be so applicable to patients who are infected with other diseases. And the dependent variable has a skewed distribution and a long tail,75% of LOS is below 11 days, while 25% of LOS ranges from 12days to 120days.

2. 50.88% of samples are over 70 years old and 83% are over 50 years old, which means that the prediction result may not be so applicable to patients who are early middle-aged or juveniles.

3. Gender distribution is relatively even in this dataset. With respect to race,60% of samples are white people ,17% are black people. Regarding ethnicity, 77% of samples are Hispanic.

Answer to Q2:

There are 18 input x variables and 1 y dependent variable in my model. I drop 4 variables from original dataset, because "Discharge. Year", "CCSR.Diagnosis.Code", "CCSR.Diagnosis.Description" are same in all samples, the content of "APR.DRG.Code" variable is same as "APR.DRG.Description".

I drop some observations that is unknown or de-identification, so about 62k observations are dropped, remaining 21,881 observations. I also dropped 4,991 observations that has over 11days of LOS(the 3rd quantile) to improve the accuracy of my model because high skewness of samples can have a huge implication on model's error, remaining 16,890 observations

```
> colnames(dt)
 [1] "Hospital.Service.Area"                "Age.Group"
 [3] "Gender"                               "Race"
 [5] "Ethnicity"                            "Length.of.Stay"
 [7] "Type.of.Admission"                    "Patient.Disposition"
 [9] "APR.DRG.Description"                  "APR.Severity.of.Illness.Code"
[11] "APR.Severity.of.Illness.Description"  "APR.Risk.of.Mortality"
[13] "APR.Medical.Surgical.Description"     "Payment.Typology.1"
[15] "Payment.Typology.2"                   "Payment.Typology.3"
[17] "Emergency.Department.Indicator"       "Total.Charges"
[19] "Total.Costs"
```

```
cols_to_drop <- c("Discharge.Year", "CCSR.Diagnosis.Code", "CCSR.Diagnosis.Description"
                  , "APR.DRG.Code")
dt <- data1[, !cols_to_drop, with = FALSE]
```

```
trainset <- dt[train == T & APR.DRG.Description != "PANCREAS TRANSPLANT"
               & Hospital.Service.Area!=""
               &Patient.Disposition!="Expired"
               &Patient.Disposition!="Another Type Not Listed"
               &APR.DRG.Description!="UNGROUPABLE"
               &Ethnicity!="Unknown"
               &Length.of.Stay<=11]

testset <- dt[train == F & APR.DRG.Description != "PANCREAS TRANSPLANT"
              & Hospital.Service.Area!=""
              &Patient.Disposition!="Expired"
              &Patient.Disposition!="Another Type Not Listed"
              &APR.DRG.Description!="UNGROUPABLE"
              &Ethnicity!="Unknown"
              &Length.of.Stay<=11]#factor APR.DRG.Description has new levels PANCREAS TRANSPLANT ar
```

testset	5037 obs. of 19 variables
trainset	11853 obs. of 19 variables

Answer to Q3:

Model	Complexity	Testset RMSE
Stepwise Linear Regression	14	1.84
CART	40	1.68
Random Forest	ntree=500&RSF size=9	1.5589

```
> summary(lr)

Call:
lm(formula = Length.of.Stay ~ Hospital.Service.Area + Age.Group +
    Gender + Race + Ethnicity + Type.of.Admission + Patient.Disposition +
    APR.DRG.Description + APR.Severity.of.Illness.Code + Payment.Typology.1 +
    Payment.Typology.3 + Emergency.Department.Indicator + Total.Charges +
    Total.Costs, data = trainset)

Residuals:
    Min      1Q  Median      3Q     Max
-73.697  -2.271  -0.481   1.355  92.991

Coefficients:
```

```
> terminal_nodes <- sum(cart.opt$frame$var == "<leaf>")
> print(terminal_nodes)
[1] 40
```

```
Call:
 randomForest(formula = Length.of.Stay ~ ., data = trainset, mtry = 9)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 9
```

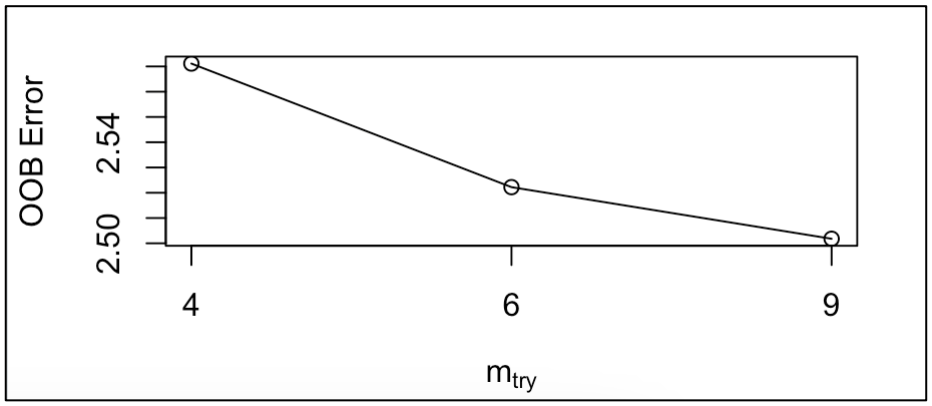
RMSE.test.cart	1.68
RMSE.test.lr	1.84
RMSE.test.RF	1.563
RMSE.test.RF.final	1.5589

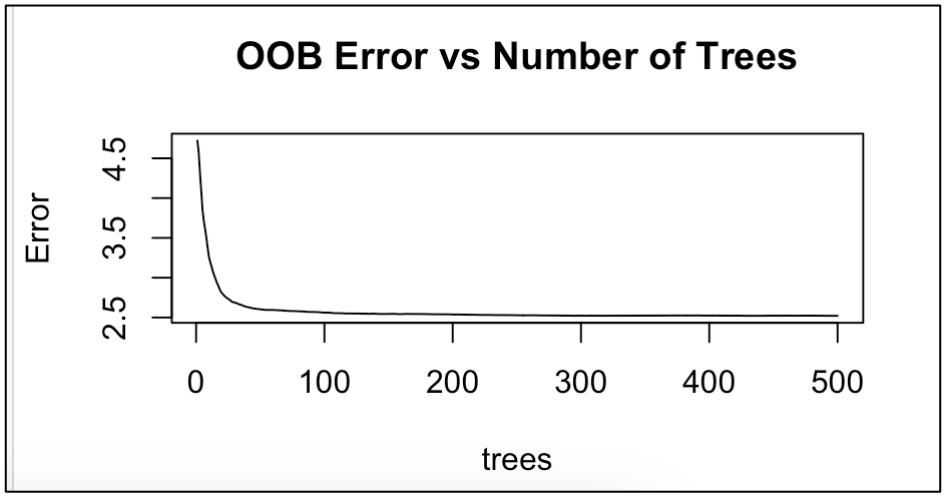
Answer to Q4:

Linear regression model: As the dependent variable has a skewed distribution, I made a log transformation to y variable to see whether it can produce a lower RMSE for linear regression model without filtering out observations of LOS over 11days. But unfortunately, log y model gave an even huge RMSE result, reaching over 180 without filtering out LOS over 11days. And Lasso regression also produce a similar RMSE as stepwise linear regression.

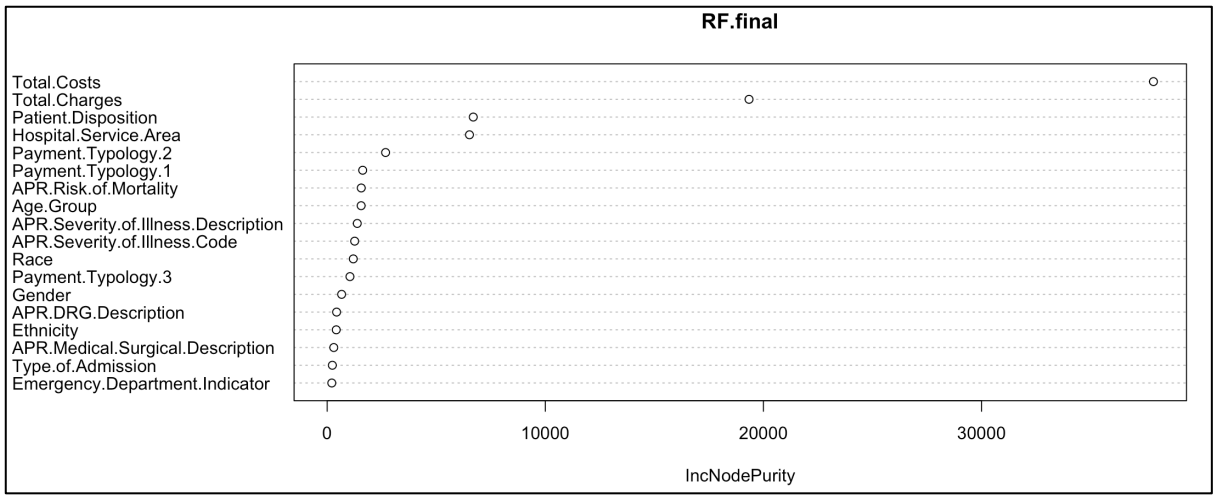
CART: I have searched the optimal CP value for my decision tree model, but the RMSE is still higher than my random forest model.

RF model: The random forest model has the lowest RMSE, which means that it has the best predictive ability among three models. And I utilized tuneRF() function (reference from ChatGPT) to search the optimal mtry and further reduce RMSE of my RF model. The default setting of mtry is 6 (because there are 18 x variables), and RMSE is 1.563. When mtry is 9, the model reflects the minimum OOB error and RMSE reduces to 1.5589. And I plotted a graph about OOB error and ntree numbers. When ntree is 500,OOB error remain steady .So there is no need to increase the tree number.





The significance of variable plot shows the top 5 significant variables. However, when I reduce the input x variables to only these top 5 variables. RMSE grows even higher to 1.64




```
> RF2 <- randomForest(Length.of.Stay ~ Total.Charges+Total.Costs+Hospital.Service.Area+APR.DRG.Descriptio
n+Patient.Disposition, data=trainset)
> RF2.yhat <- predict(RF2, newdata = testset)
> RMSE.test.RF2 <- round(sqrt(mean((testset$Length.of.Stay - RF2.yhat)^2)),2)
> print(RMSE.test.RF2)
[1] 1.64
```

This random forest model may only have high predictive ability for those Septicemia patients whose total cost of cure is less than 240,000 and total charge less than 1,000,000. When hospital admitted a Septicemia patient and input the estimated cost and charge of cure and other detailed information about the patient, the model can tell a relatively accurate length of stay. However, when comes to a more severe Septicemia patient, prediction may not be so accurate.

This model based on the dataset of patients who are all discharged in 2022, so it do not consider the fluctuation caused by year.

Answer to Q5:

We can extend the model by collecting more data from more severe patients who have LOS more than 11days, or who are infected with other diseases, or who are discharged in other year rather than 2022. In that case, we can build a model that have a wider business application.