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

Full Name	Raashi Singh
Email Address	Raashi001@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

Answer to Q1: 2

Answer to Q2: 4

Answer to Q3: 5

Answer to Q4: 7

Answer to Q5: 9

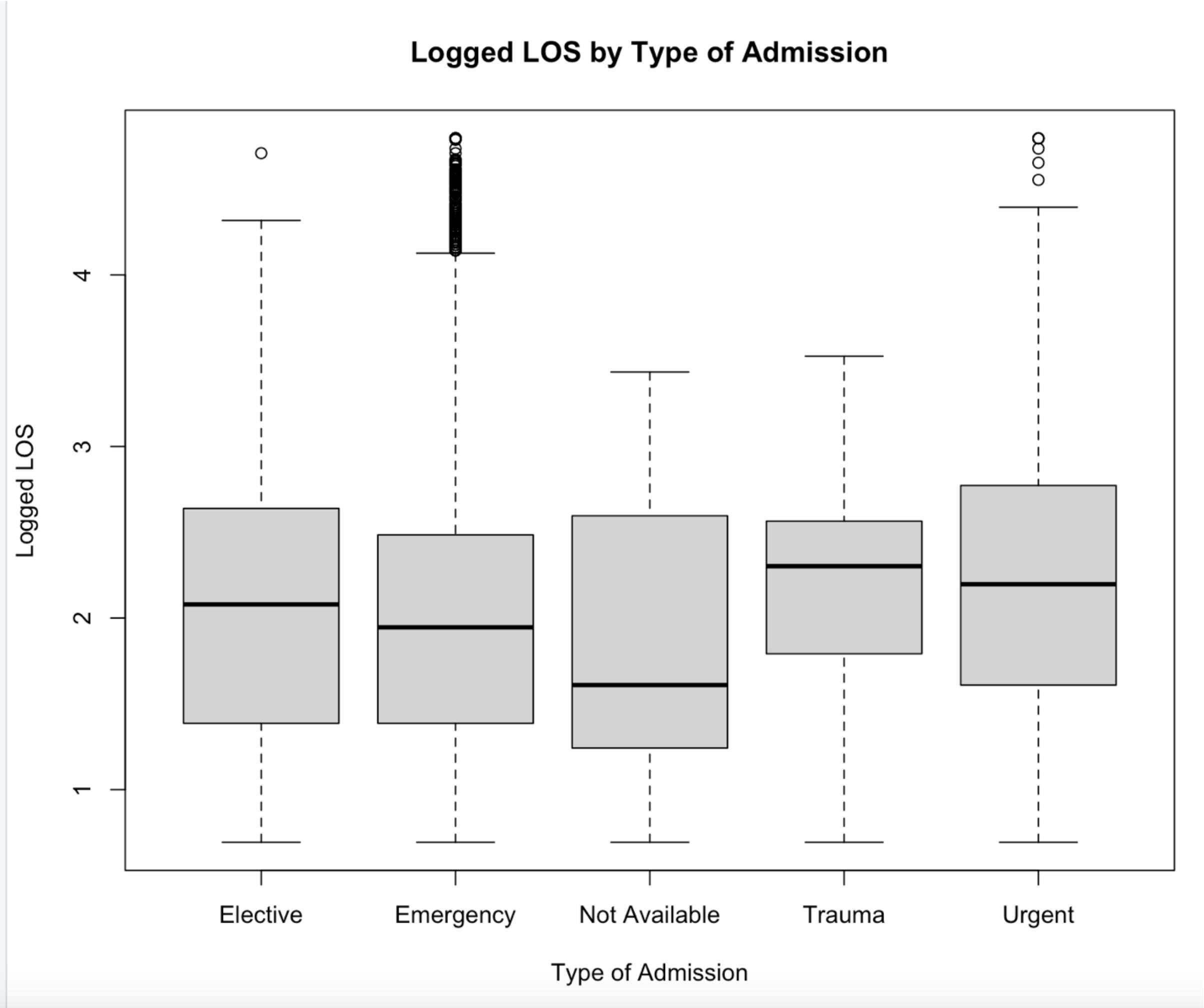
For each question, please start your answer in a new page.

Answer to Q1:

For my first notable finding, I wanted to find the correlation between the APR Severity of Illness and the Length of Stay, and I got the value of 0.3832153, which depicts a positive weak correlation between a patient's severity of illness and length of stay, which is surprising as my initial thoughts about this correlation was that Severity of Illness, the extent to which a patient is unwell, would be a stronger attribute to how long the patient stays in the hospital. However, this just tells me that there are stronger predictor variables than Severity of Illness in the dataset, for Length of Stay.

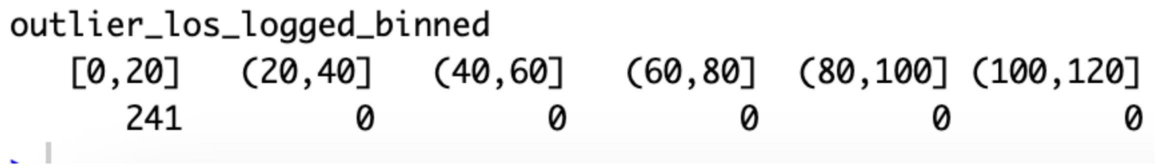
For my second notable finding, I performed a box plot between Type of Admission and Length of Stay, where I saw that patients who have come in due to a trauma injury have a higher median of LOS compared to patients who come under urgent or elective situations. This may be due to patients under trauma having a serious (or possibly, life threatening) injury (like head trauma), which results in them taking a longer recovery time, so higher LOS.

The boxplot is as follows: (LOS is logged. Reason is given in common note of qn 3)



For my third notable finding, I wanted to see what the LOS for patients is, for patients who are OUTLIERS for total charge. (I have picked total charge here, as total charge is the actual sum that patients are paying to the hospital, at least this is what I interpreted from the Data Dictionary).

These are the results I see.



```
outlier_los_logged_binned
   [0,20]   (20,40]   (40,60]   (60,80]  (80,100] (100,120]
      241         0         0         0         0         0
```

I have represented it in a table because this is a neater way to illustrate the info. I ended up logging my Total Charges because I discovered it to have an incredible right skewness. I make here an interesting discovery. I see that 241 patients who are outliers in (log of) total charges, all have a LOS of 0-20 days. This is interesting as one might think that if a patient is paying a high amount of charge, it may be due to them having a high length of stay in the hospital (which would also include higher costs of doctor's and nurse's care for the longer LOS). So it is interesting that outliers of TC who are paying high amounts of money to the hospital, are staying for a maximum of 20 days. This could mean that maybe they are undergoing an expensive procedure, which has a quick recovery rate.

Answer to Q2:

- Age Group (The older the patient, the longer their recovery time so longer their LOS)
- Gender
- Type of Admission (as seen in Qn 1, patients under emergency have a higher LOS)
- APR Severity of Illness Code (the more severe a patient's illness, the longer their LOS)
- Emergency Department Indicator (similar justification to type of admission)
- APR DRG Description (as this is the main description of the patient's illness, and hence should be strongly correlated with los)
- Race
- Payment Typology 1 (payment methods may affect the LOS as patients who are insured may be more inclined to have a higher LOS as they are convinced that the costs will be taken care by their insurance, but people who may have to take the burden of self-paying for their medical care may want to leave the hospital as soon as possible for lowest costs incurred)
- Payment Typology 2 (same as Payment Typology 1)
- Payment Typology 3 (same as Payment Typology 1)

With the new dataset, we have 28108 obs. of 12 variables (10 potential x variables with LOS and logged LOS (please look at the common note in qn 3 for explanation of logged LOS) as well).

Answer to Q3:

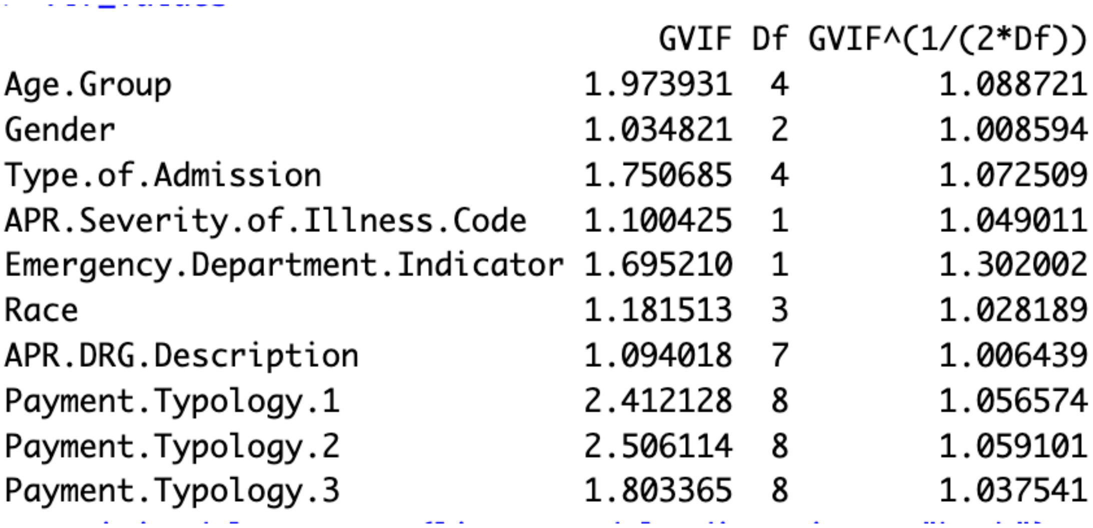
Model	Complexity	Testset RMSE (LOS was logged)
Linear Regression	8	0.6553759
CART	6	0.6632084

For both (common note):

I logged my Length of Stay in my data cleaning, as I realised that my LOS is very right skewed. Hence, I logged LOS. Thus, my RMSE values have LOS as logged, and may be different from other student's RMSE values, or the typical RMSE values we are supposed to obtain.

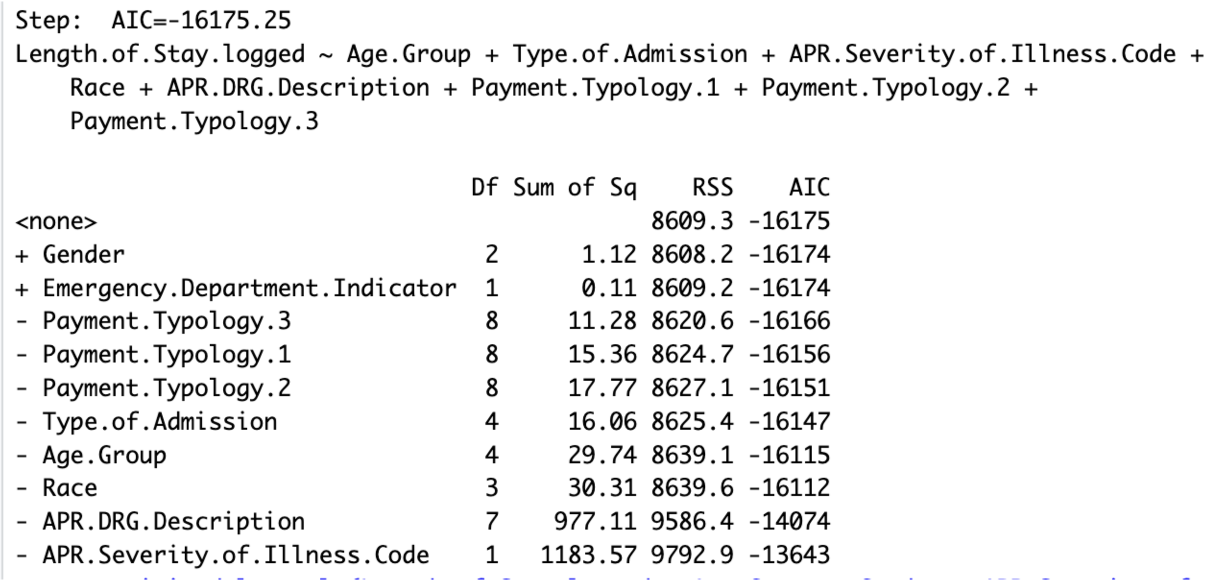
For Linear Regression:

To optimise the models, I first checked for multicollinearity between the variables. Since all the gvif values were below 10, this did not help optimise my model.



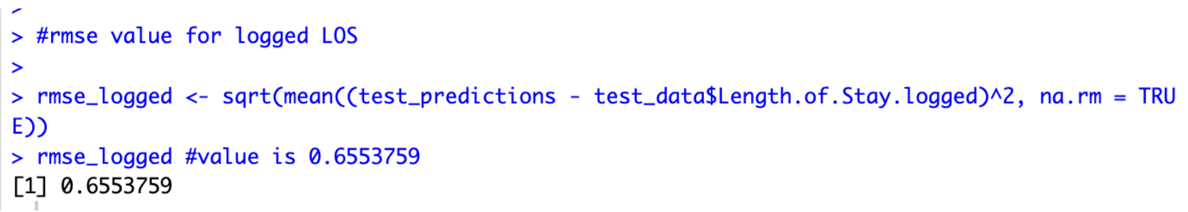
```
                                GVIF Df GVIF^(1/(2*Df))
Age.Group                   1.973931  4        1.088721
Gender                      1.034821  2        1.008594
Type.of.Admission           1.750685  4        1.072509
APR.Severity.of.Illness.Code 1.100425 1        1.049011
Emergency.Department.Indicator 1.695210 1      1.302002
Race                        1.181513  3        1.028189
APR.DRG.Description         1.094018  7        1.006439
Payment.Typology.1          2.412128  8        1.056574
Payment.Typology.2          2.506114  8        1.059101
Payment.Typology.3          1.803365  8        1.037541
```

Then, to further optimise the model, I used stepwise regression (with AIC), and the model suggested to remove Gender and Emergency Department Indicator. I determined this by looking at the model with the lowest AIC (-16175.25). So, my new optimised linear model excludes Gender and Emergency Department Indicator. And hence, I have 8 x predictor variables.



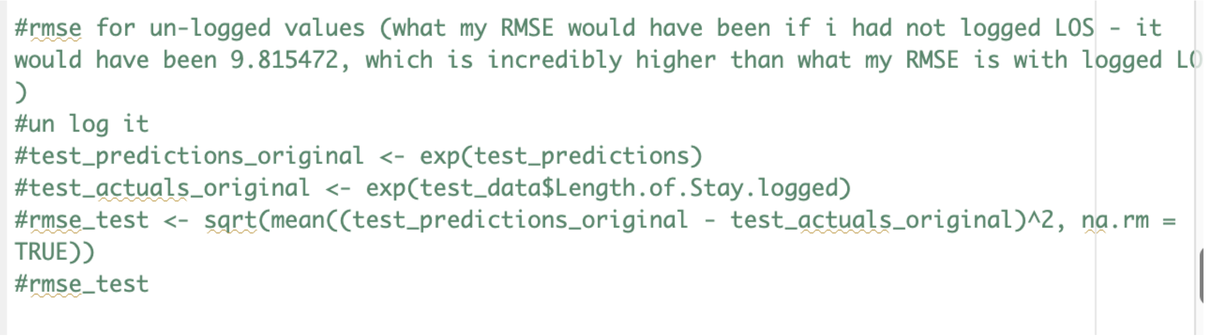
```
Step:  AIC=-16175.25
Length.of.Stay.logged ~ Age.Group + Type.of.Admission + APR.Severity.of.Illness.Code +
    Race + APR.DRG.Description + Payment.Typology.1 + Payment.Typology.2 +
    Payment.Typology.3

                                 Df Sum of Sq    RSS    AIC
<none>                                        8609.3 -16175
+ Gender                          2      1.12 8608.2 -16174
+ Emergency.Department.Indicator  1      0.11 8609.2 -16174
- Payment.Typology.3              8     11.28 8620.6 -16166
- Payment.Typology.1              8     15.36 8624.7 -16156
- Payment.Typology.2              8     17.77 8627.1 -16151
- Type.of.Admission               4     16.06 8625.4 -16147
- Age.Group                       4     29.74 8639.1 -16115
- Race                            3     30.31 8639.6 -16112
- APR.DRG.Description             7    977.11 9586.4 -14074
- APR.Severity.of.Illness.Code    1   1183.57 9792.9 -13643
```

Then, I predict on my test set and obtain a RMSE value of 0.6553759.



```
> #rmse value for logged LOS
>
> rmse_logged <- sqrt(mean((test_predictions - test_data$Length.of.Stay.logged)^2, na.rm = TRU
E))
> rmse_logged #value is 0.6553759
[1] 0.6553759
```

My RMSE value would be a lot higher (around 9.81) if LOS was not taken to be logged! Hence, further justifying why my LOS was taken as logged LOS.



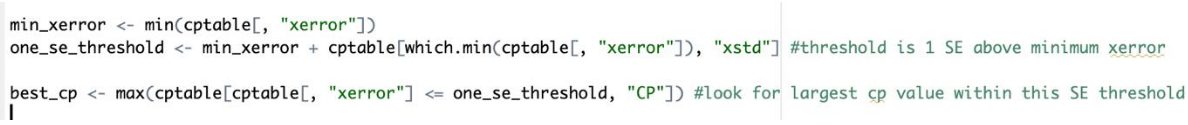
```
#rmse for un-logged values (what my RMSE would have been if i had not logged LOS - it
would have been 9.815472, which is incredibly higher than what my RMSE is with logged LO
)
#un log it
#test_predictions_original <- exp(test_predictions)
#test_actuals_original <- exp(test_data$Length.of.Stay.logged)
#rmse_test <- sqrt(mean((test_predictions_original - test_actuals_original)^2, na.rm =
TRUE))
#rmse_test
```

For cart:

LOS also kept as logged here, as if it is not logged and the model is run with the right-skewed LOS, the RMSE value would be around 9.28 (a LOT higher than what my RMSE value is, when LOS is logged).

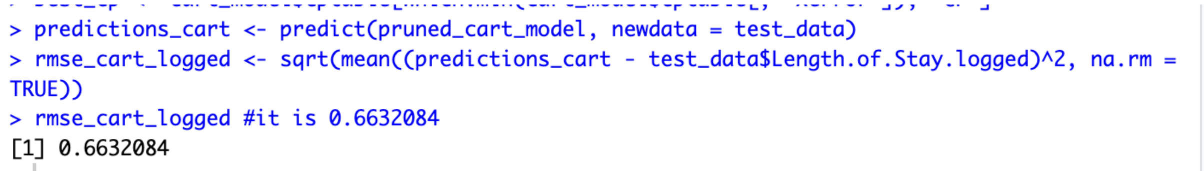
Cart model is initially run with all predictor x variables from Question 2.

Then, I can see the pruning sequence and the 10 fold errors in a table form, by the printcp command (like we have done in class). After which, I look at what the best cp can be. Initially, I had found the best cp using the min cv error method, then realised that it is an unstable solution, as a small change in data could lead to a different solution (as taught to us in the lecture). So I approached the 1 SE rule. So, I found the minimum xerror, found the 1 SE error threshold, and then found the best cp, by looking for the largest cp value within that 1 SE threshold. Here are the lines of code I wrote out for this:



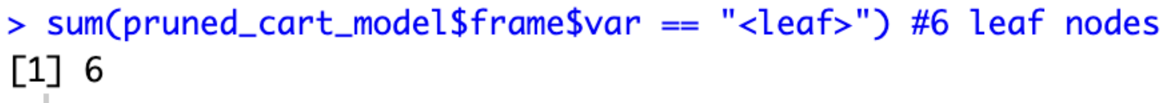
```
min_xerror <- min(cptable[, "xerror"])
one_se_threshold <- min_xerror + cptable[which.min(cptable[, "xerror"]), "xstd"] #threshold is 1 SE above minimum xerror

best_cp <- max(cptable[cptable[, "xerror"] <= one_se_threshold, "CP"]) #look for largest cp value within this SE threshold
```

I then prune the tree according to this best_cp value that I have calculated, after which I find my RMSE value.



```
> predictions_cart <- predict(pruned_cart_model, newdata = test_data)
> rmse_cart_logged <- sqrt(mean((predictions_cart - test_data$Length.of.Stay.logged)^2, na.rm =
TRUE))
> rmse_cart_logged #it is 0.6632084
[1] 0.6632084
```

Lastly, I find the number of leaf nodes of my final optimised CART model!



```
> sum(pruned_cart_model$frame$var == "<leaf>") #6 leaf nodes
[1] 6
```

Answer to Q4:

My Linear Regression and CART models work similarly in performance, with similar test set RMSE values of 0.6553759 (Linear Regression - LR) and 0.6632084 (CART). As aforementioned, these values are lower than the typical RMSE values other students may be getting, as I realised that Length of Stay (LOS) was incredibly right-skewed. Hence, I logged the LOS and used the logged LOS in my 2 models, improving the accuracy of these models by lowering their test set RMSE values. It is important to note, that though very close to one another, my RMSE value for CART is slightly higher than for LR, which could indicate that my variables relationships with one another in the dataset may be better described by a linear relationship, which a CART may not be able to represent.

Furthermore, the complexity values of 8 in LR and 6 in CART, suggest a relatively simple model, which would hence be easier to be read and understood by stakeholders, like hospital staff and nurses. Due to the lower complexity, there is also a lower chance of overfitting in my CART model, which means it is a more generalised model, which can also work even if my dataset changes. This is essentially important in hospital datasets, which are ever changing, with new patient records coming in everyday, at a huge rate.

By understanding which specific variables are significant to predict LOS, we can understand better which patient is more likely to have a longer LOS, and hospitals can prepare better beds and rooms for such patients. Therefore, there can be better resource management, where hospitals can allocate more suitable rooms to patients depending on their LOS. Furthermore, knowing when patients are ready to be discharged, also helps the hospital plan better to provide the patients with a more streamlined and efficient discharge. If hospitals can also understand what factors or predictor x variables result in higher LOSs, hospitals can devise better healthcare plans for them, to hopefully shorten their LOS, which can result in hospitals saving on operational costs. Better healthcare plans can mean better healthcare aid etc. Lastly, better resource planning would also greatly improve patient satisfaction, and reduce patient anxiety, as patients are more informed on how long they would be staying at the hospital for, and can hence plan their own lives better, for instance letting their bosses know when they would be back on the job etc.

It was mentioned in the CBA Question Paper that currently, the estimation of a patient's LOS is done using the mean LOS in each CCSR diagnosis. However, this can prove to be a very limited diagnosis, since the only predictor x variable here is CCSR diagnosis. With more predictor x variables being included in the LR and CART models, the predicted LOS of a patient can be more accurate. Furthermore, looking at mean LOS, as is currently done, may give inaccurate results, since LOS is a highly right-skewed data, and mean takes into account all outliers as well.

Answer to Q5:

As for improvements for my models, I could have tried to optimise them even further, to come up with more accurate models. For linear regression, I could have looked at interaction terms, which helps to understand how the relationship of one predictor variable on the response variable changes, with regards to the level of another predictor variable. This would be easier to understand with an example. If we are looking at 2 predictor variables: Age and the Severity of Illness separately, we would understand that a higher value of both predictor variables, result in higher LOS. However, it is interesting to understand, with the help of interaction terms, that severity of illness would have a greater impact on LOS if the patient were older. Hence, this takes the relationships between such predictor variables a step further and looks at them with a more nuanced view. This also results in a better LOS prediction of the patient. Patients who are determined to be high-risk patients with a specific combination of variables can also be targeted and looked after properly from the beginning. This kind of better planning can also help the hospital save unnecessary medical expenditures later on.

As for my CART model, I could have carried out Stratified Sampling. This is a sampling technique where each category in a column of a dataset is represented proportionally in the sample. This is crucial in cases where a column may be greatly imbalanced. For instance, a column I included in both optimised models, Type of Admission, is a very imbalanced data. It has around 27000 emergency cases, with the rest of the 4 categories having a sum total of around 1000 cases. Hence, stratified sampling would have helped give a better representation of the categories. This significantly improves the prediction of LOS. This can be incredibly important in terms of race and other social variables, where a particular race may be under-represented if not for Stratified sampling. Thus, this sampling technique makes it a more inclusive model, where patients' satisfaction is also increased due to this inclusivity.

These models can also be made to be more dynamic in nature. Any medical results, or reports can be automatically inputted into the models, to predict the patient's LOS better. For instance, if a patient checks in relatively healthier, the predicted LOS may be a few days. However, if the patient's condition gets worse and if the models are not dynamic, the predicted LOS would be lower than the actual LOS, resulting in an inaccurate model.

Furthermore, the dataset can be more comprehensive. For instance, pre-existing medical issues of patients can be inputted into the models as well, to see how they may affect their current condition, and if that might potentially raise the LOS. In addition to pre-existing medical issues, a patient's current medicine list could also be taken into account, because they give a more comprehensive picture of a patient's health. To further create a more inclusive environment, a patient's first language can also be taken into account, so that nurses and doctors understand how to deal with this patient, and make them feel more comfortable. Another social factor to be considered, can be a patient's emergency contact, or social network, to better understand who to liaise with, lest some medical complications arise.

Lastly, more suggestions can be taken in from stakeholders such as doctors and nurses, who better understand and work everyday alongside such patients. Their inputs would make this model increasingly comprehensive and more accurate to predict LOS for patients.