**Project Report**

**Introduction: -**

Care One of the hospitals is a medium-sized teaching hospital (about 350 beds) affiliated with a large university's medical school. It has long been regarded as a leader in providing high-quality care and has recently been awarded magnet status. The management of Care One Hospital have recently expressed concern over an increase in the duration of stay of patients at their facility. Following an investigation into this issue, multiple Care One Hospital providers discovered that the time it takes to acquire a requested X-Ray is unnecessarily long.

The key is to employ regression to aid Care One Hospital in their radiology process improvement program. Interpret the variables in the final model(s) and consider how the hospital could use them to improve its process.

**Pre-treatment of data: -**

* Patients, physicians, and radiology technicians have all been allocated a unique identity number based on the de-identified data. We have removed the unique identifier as a predictor variable because it has nothing to do with the model which we going to predict. The below command is used to remove the unique identifier from the data.

**Data<- subset(Data, select = -c(Unique.Identifier)).**

* We converted strings as factors because data is stored more efficiently, because each unique string gets a number and whenever it's used in your data frame you can store its numerical value (which is much smaller in size). The command which we used to change the strings to factors is written below.

**Data <- read.csv(file = 'C:/Users/Saipa/DownloadsRadDat\_5344.csv', stringsAsFactors = TRUE)**

* There are some none values in the data, so we used this command to remove the none values.

**Data<-na.omit(Data)**

**Model Adequacy of Basic Model(Model): -**

Regressing ordered to complete x-rays with all the predicted variables.

1. **Residual vs Fitted: -**

Diagram

Description automatically generated

* In residual vs Fitted graph there is no pattern seen the number of points are suffiecient to tell the model is adequate.

1. **Normal to Quantile to Quantile: -**

Diagram

Description automatically generated

* By checking the model adequacy, the Normal Quantile to Quantile probability looks ok because it is real time data.
* So, we can check by using Box-Cox function

Chart

Description automatically generated

* By seeing the graph, the lambda value is close to the interval zero, so the data transformation of data is no need to be done.

1. **Standardized Residual vs Scale location Graph: -**

Diagram

Description automatically generated

1. **Residuals vs Leverage: -**

Chart

Description automatically generated

* According to the graph there is no point which is greater than 1. so, Cooks distance Di<1 for all the points. so, all the points are not influential points.
* So, we don’t need to remove any points on the data.

**Summary of The Basic Model: -**

Residual standard error: 6268 on 43358 degrees of freedom

Multiple R-squared: 0.2488, Adjusted R-squared: 0.2459

F-statistic: 87.55 on 164 and 43358 DF, p-value: < 2.2e-16

* Even though the r square value is very small by checking the p-value we can proceed with the model because the p-value is very small.

**Reduced Model: -**

**Model1: -**

* We are removing in rad room and patient’s age

Justification For Patient’s Age: -

* From the summary of the basic model, we got patient age as

Estimate Std. Error t value Pr(>|t|)

PatientAge -1.163 1.544 -0.753 0.451273

* We got beta as negative i.e.., time taken to complete order is inversely proportional to age, but it doesn’t make any sense, and the time to complete order doesn’t depend upon the age and the term is also not significant from the model.

Justification For in Radiation Room: -

* After regressing we got in Rad room parameters as below,

In.Rad.Room NA NA NA NA

* There are no parameters shown and nothing is changed in the model after removing it and the term is also not significant.

Parameters of Model1 (R-square value and p-value) is nothing but the basic model after removing these insignificant terms. So, we are going to proceed further to model2.

**Model2: -**

* Now we have checked for multi collinearity of model1 for removing further insignificant terms

vif(Model1)

GVIF Df GVIF^(1/(2\*Df))

Radiology.Technician 1.020783 1 1.010338

CatalogCode 24.485289 120 1.013414

PatientTypeMnemonic 2.215225 3 1.141746

Priority 3.111434 1 1.763926

Loc.At.Exam.Complete 17.894700 24 1.061936

Exam.Completed.Bucket 1.482583 2 1.103455

Exam.Room 163.316249 12 1.236544

* The variance inflation factor for the Exam.Room has more, and it should not be removed because it is the most significant term in the model, So, it is okay f it has more co-linearity.
* The next term which has more Multi co-linearity (VIF) is Catalog-code I.e., why it should be removed, and it is also not significant term in the model.

Our model2 is in ordered to complete x-ray regressed with all factors in model1 and excluding Catalog-code.

Residual standard error: 6310 on 43479 degrees of freedom

Multiple R-squared: 0.2364, Adjusted R-squared: 0.2357

F-statistic: 313 on 43 and 43479 DF, p-value: < 2.2e-16

The above parameters are the summary of our model 2

* These looks to be similar as our model1 but, there is slight difference in our R-square value.

vif(Model2)

GVIF Df GVIF^(1/(2\*Df))

Radiology.Technician 1.017511 1 1.008717

PatientTypeMnemonic 1.932612 3 1.116068

Priority 3.085613 1 1.756591

Loc.At.Exam.Complete 14.195080 24 1.056824

Exam.Completed.Bucket 1.443386 2 1.096088

Exam.Room 12.797519 12 1.112065

* The above is the VIF of our Model 2. According to the rule of thumb the highest vif values needs to be removed those are Loc.At.Exam.Complete and Exam.Room ,but these terms cant be removed because these are the most significant terms in the model.
* The above predictors are our final predictors and our model 2 is the final model

**Backward direction stepwise regression (Model B): -**

* In backward we start with the full model and reduce the parameters or drop the variables the variables do not predict anything on the dependent measure or removed one by one.
* We got backward step wise regression formula as

Call:

lm(formula = Ordered.to.Complete...Mins ~ CatalogCode + PatientTypeMnemonic + Priority + Loc.At.Exam.Complete + Exam.Completed.Bucket + Exam.Room, data = Data)

* From the above formula we need to remove the Catlog-code,when we remove the Catlog -code the formula reduces to

Call:

lm(formula = Ordered.to.Complete...Mins ~ PatientTypeMnemonic + Priority + Loc.At.Exam.Complete + Exam.Completed.Bucket + Exam.Room, data = Data)

* The above regression model is the best regression model according to Backward direction stepwise regression.

**VIF of Backward Regression Model: -**

GVIF Df GVIF^(1/(2\*Df))

PatientTypeMnemonic 1.932406 3 1.116048

Priority 3.085613 1 1.756591

Loc.At.Exam.Complete 14.013914 24 1.056542

Exam.Completed.Bucket 1.438855 2 1.095227

Exam.Room 12.781088 12 1.112006

**Forward direction stepwise regression (Model F): -**

* In forward we start from null and add predictors to the model this is repeated with the variable and predicts the most on the dependent measure this procedure continues until doesn’t not add anything to the prediction model anymore.
* We got formula for forward stepwise regression is,

Call:

lm(formula = Ordered.to.Complete...Mins ~ Loc.At.Exam.Complete +Priority + In.Rad.Room + PatientTypeMnemonic, data = Data)

* The term In.Rad.Room must be removed because it doesn’t contributes to the model.
* From the above formula we need to remove the In.Rad.Room,when we remove the In.Rad.Room the formula reduces to

Call:

lm(formula = Ordered.to.Complete...Mins ~ Loc.At.Exam.Complete +Priority + PatientTypeMnemonic, data = Data)

**VIF of Forward Regression Model: -**

GVIF Df GVIF^(1/(2\*Df))

Loc.At.Exam.Complete 3.824592 24 1.028341

Priority 2.833871 1 1.683410

PatientTypeMnemonic 1.919481 3 1.114801

**All Models- Best Sub-Set Model**

* The best subsets regression method is a model selection method that involves evaluating every conceivable combination of predictor variables and then picking the best model based on statistical criteria.
* We are considering predictor variables as terms which are present in our final model which is model 2 i.e., the terms excluding patient’s age, Catlog-code and In.Rad.Room.

A picture containing table

Description automatically generated

Table

Description automatically generated

1. Allowing us to identify the best overall model, where best is defined as the model that maximize the adjusted R2 and minimize the prediction error (RSS, cp and BIC).
2. The adjusted R2 represents the proportion of variation, in the outcome, that are explained by the variation in predictors values. the higher the adjusted R2, the better the model. The model with the lowest AIC is selected.

* From the above two points the model 5 and model 6 are the best models according to the table and one of the models is our model 2.

1. Model 6 from the above best model sub-set table is our model 2 which is our final model

**Parameter’s Graphs of Our Best Model Sub-Sets: -**

Chart, line chart

Description automatically generated

* The above and below graph shows R-Square, Adjusted R-Square , AIC ,Sbc,Bic of different models in best sub-sets regression table.

Chart, line chart

Description automatically generated

**Interpretations: -**

* In our final model (model 2) we got almost all terms are significant and the final model p-value is 2.2e^-16 i.e., the model is significant and having intercept as high-value and positive.
* VIF’S of our final model looks decent and some terms has more VIF values, but These are the most significant terms in our model, so we are not removing them.

We are not removing terms like

1. Radiology. Technician

2. Patient type Mnemonic

Because in real time data analysis these are the most important terms which predict our model, even though they are not significant terms in model we must keep them and that’s why we are not removing them.

* The most significant terms in the model are Exam. Completed.Bucket,prioritystat and Exam.Room
* From 12:00am to 8:00 am that is night shifts average time taken to complete the order x-rays is 5067 min which is a lot of time so the staff must concentrate on the night shifts, and they need to have a surveillance and needs to know what is going wrong.
* For routine check-up the time taken for the x-rays takes much time than stat check-up

Suggestion- The staff must concentrate more on routine check-up.

* For IP patient type mnemonic, the average time taken to complete the x-rays is 3025 min and for other type mnemonic there is short type of time span so the staff must concentrate on Ip type mnemonic.
* The Dx chest Portable X-rays it is taking more time than other X-rays.
* The average time taken for getting x-ray in these rooms Dx Portable, Dx Rm 1, Dx Rm 2, Dx Rm 5 pan, Dx Rm 6 (EC), Dx Rm 9, is more. And staff monitor or keep a regular check on these rooms and find what is the cause of delay in the x-rays in these rooms.
* At These locations the average time taken is more than other rooms 3W, 4E, 4W, 5E, 5W, BICU, BIMC, CHCU, CICU, GTU, MICUN, MICUS, PEDA, PICU, SCU, SICU, and staff monitor or keep a regular check on these rooms and find what is the cause of delay in the x-rays in these rooms.