Package Pricing at Mission Hospital

Adithi Mahesh

Anika Gupta

Archit Jaiswal

Suhas Kotaki

Data Exploration

Total 248 observations across 24 (raw data) / 51 (modified data) columns

Target variable – Ln(TOTAL COST TO HOSPITAL)

Technology Acceptance Model

Personal Data	Health Variables	Stay Variables	Admission time Statistics
Age_category Gender BMI = Height / Weight Marital Status	Key complaint Codes Past Medical History Code Implant	Total Length of Stay Length of Stay-ICU Length of Stay-Ward	HR Pulse BP- High / Low Respiratory Rate HB, Urea, Creatinine Mode, State, Admission

Target Variable : Total Cost to Hospital

Data Preparation

Removing columns

Removed columns with insufficient observations as "1"

- CAD-SVD
- CAD-VSD
- Other-general
- Confused state

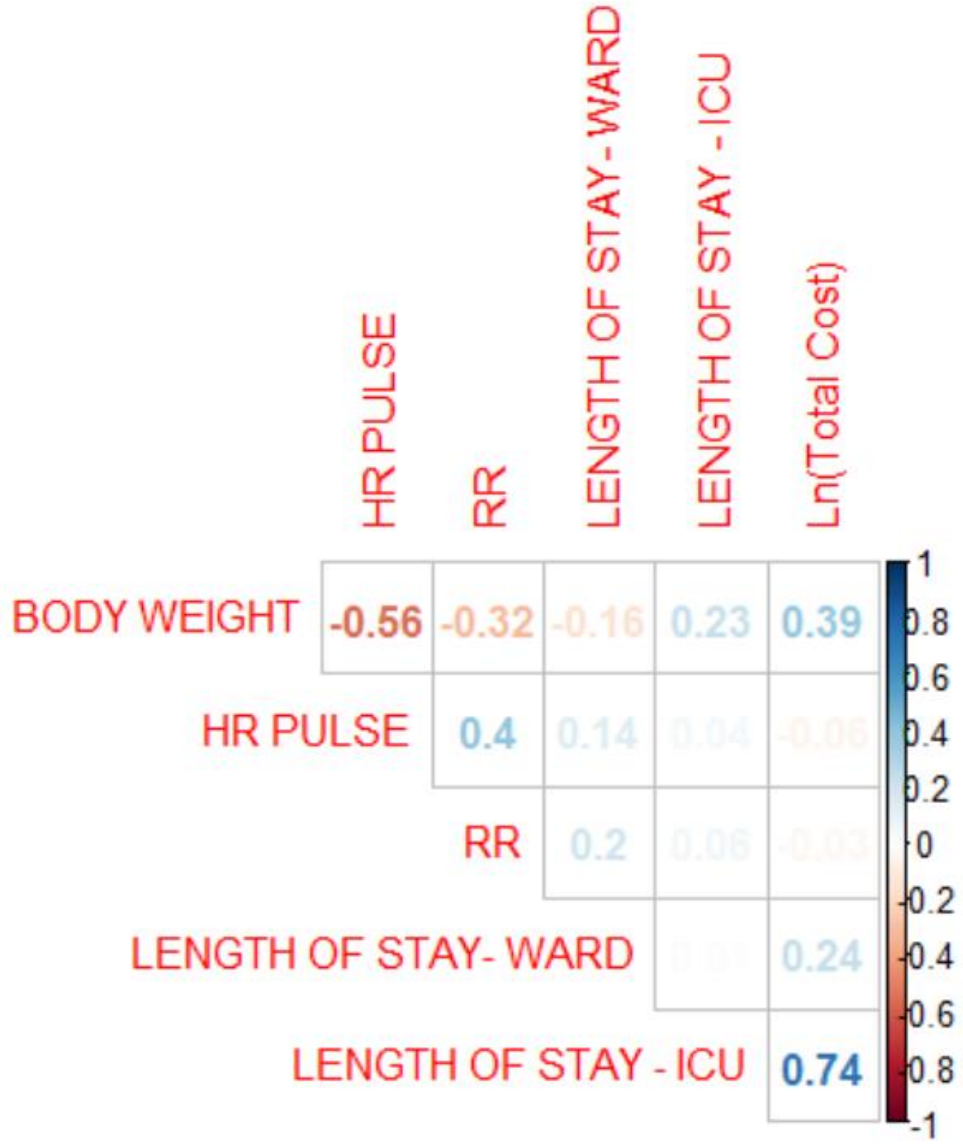
Missing Values

- Removed columns with more than 10% missing values
  - Creatinine has 13% missing values

Bivariate

- T-test
- ANOVA
- Correlation Matrix

Variable Correlation



High correlation with target variable Ln(Total Cost)

- Length of Stay - ICU is highly correlated with correlation coefficient of 0.74
- Body Weight - 0.39
- Length of Stay Ward - 0.24

Model Deployment

- **Dummy variables** were created to prepare data for regression model
  - Age-Category <10, 11-25, 26-50, 50 and above
  - BP-Ranges – Low, Normal, High, Critical
  - BMI – Underweight, Normal, Overweight, Obese
- **Feature engineering**
  - Hemoglobin normal range: Female-12 to 15.5, Men – 13 to 17.5
  - Urea normal range: 7 to 20 mg/dl
- **Null Value Handling**
  - BP Ranges – Imputed 'Normal' BP range for null values which were Juvenile Patients
  - Urea – Imputed 'Normal' Urea level for 11 null values. Assumption: Urea measurement is not critical for that patient.

T-test : Variable Selection

T-test results showed non-significance for some variables including:

- other-heart, other-nervous, other-tetralogy, Diabetes1, hypertension3, hemoglobin
- Above variables were removed

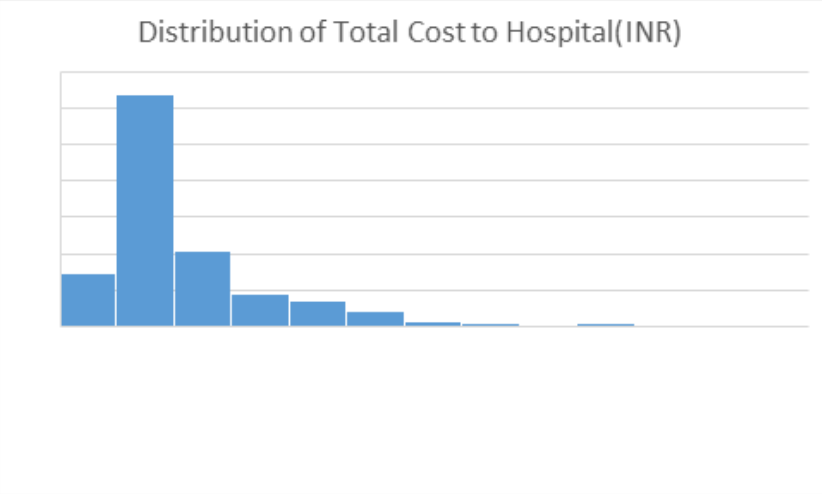
```
> unlist(c)
 [1] 0.004891094559017 0.0000000002521739 0.0013245835972240 0.000063921537185
 [5] 0.000458973760583 0.0000730461582812 0.3398040815335372 0.071196202392515
 [9] 0.624783992716502 0.3524801475678168 0.0084629667228695 0.000040051819940
[13] 0.183278994288462 0.0005908182800583 0.0018282991131916 0.037888495390284
[17] 0.456095273203767 0.0119428338101021 0.6207373720410457 0.101850144577240
[21] 0.000322741531034 0.0003350781592980
```

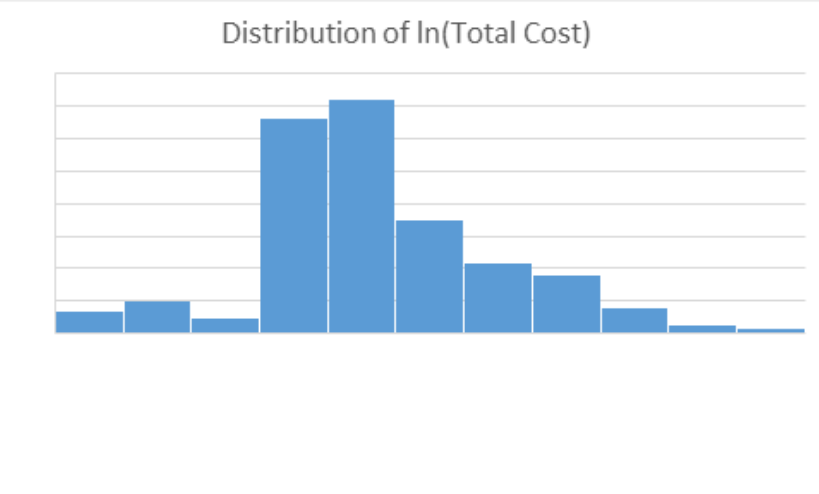
ANOVA Test

Significant categorical variables include:

- Age Categories
- BMI
- BP Ranges

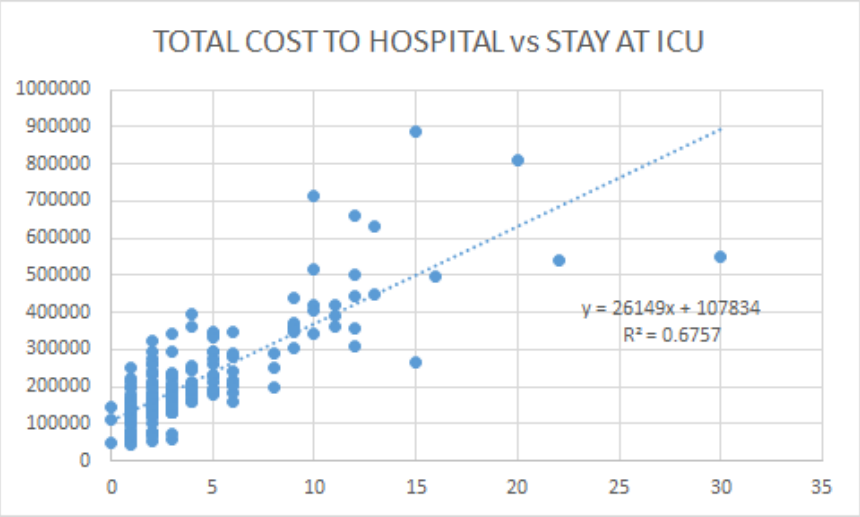
```
> summary(anova_AGE)
             Df Sum Sq Mean Sq F value          Pr(>F)
md$AGE        3  10.41   3.470   17.95 0.000000000147 ***
Residuals   244  47.18   0.193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(anova_BMI)
             Df Sum Sq Mean Sq F value     Pr(>F)
md$BMI        3   6.79  2.2620   10.86 0.00000101 ***
Residuals   244  50.80  0.2082
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(anova_BP)
               Df Sum Sq Mean Sq F value  Pr(>F)
md$`BP Range`   3   2.95  0.9827   4.388 0.00498 **
Residuals     244  54.64  0.2239
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

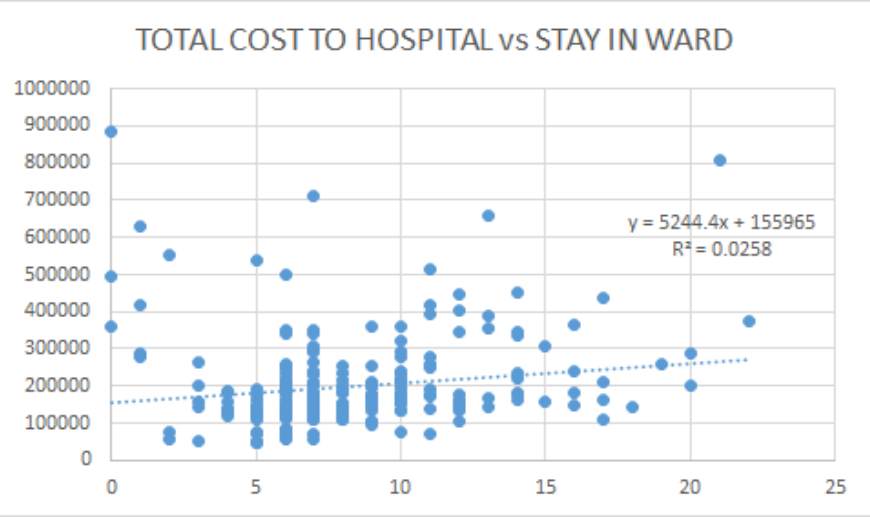




Considered ln(Total Cost) instead of Total Cost to Hospital(INR)

Our variable has a right skew. Taking the log would make the distribution of our transformed variable appear more symmetric.





Removed Stay at hospital – ICU and Ward

It is a reflective construct variable

Cannot predict length of stay accurately at the time of admission to compute total treatment cost

Variable Interaction

Several variable Interactions were significant in identifying Total cost to Hospital

Interaction	Significant (Yes/No)
Patient used Ambulance and Had impant	Yes
Age <10 and Rheumatic Heart Disease present	No
Age>50 and Coronary artery disease-TVD	Yes
Age>50 and Coronary artery disease-DVD	Yes
ACHD and Body Weight	Yes
RHD and Cost of Implant	Yes

Some variable interaction coefficients are mentioned below

```
relevel(md$AGE, ref = "2")3  0.01184    0.09698   0.122            0.902976
md$`IMPLANT USED (Y/N)`Y     0.49271    0.07426   6.635      0.000000000217 ***
md$`Amb*implant`1            0.40247    0.14700   2.738            0.006651 **
md$`juvenile*RHD`1          -0.43466    0.39690  -1.095            0.274562
md$`juvenile*ACHD`1         -0.37922    0.11988  -3.163            0.001763 **
md$`Age*TVD`1                0.37432    0.10963   3.414            0.000751 ***
md$`Age*DVD`1                0.47069    0.10775   4.368      0.000018702194 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3881 on 238 degrees of freedom
Multiple R-squared:  0.4319,    Adjusted R-squared:  0.4104
F-statistic:  20.1 on 9 and 238 DF,  p-value: < 0.00000000000000022
```

Impact of Body Weight on Total Cost to Hospital

Average Cost to hospital for 50Kg patient : INR198,723
Body Weight coefficient : 0.008434

For everyone unit increase in body weight, there will be 0.008 times ln() value of cost.
(e^0.0084) = 1.0084.
0.0084 increase from the actual value.

Likely increase in cost to hospital with increase of 1 Kg of patient weight : 198,723 * 0.0084 = INR 1669

<u>Conclusion:</u> A patient weighing 51Kgs is likely to spend INR 1,669 more than the patient weighing 50Kgs

```
> fit2<- lm(md$`Ln(Total Cost)` ~ md$`BODY WEIGHT`)
> summary(fit2)

Call:
lm(formula = md$`Ln(Total Cost)` ~ md$`BODY WEIGHT`)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35686 -0.28031 -0.02637  0.23793  1.51232

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      11.745585   0.056642 207.364  < 2e-16 ***
md$`BODY WEIGHT`  0.008434   0.001286   6.559 3.16e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4672 on 246 degrees of freedom
Multiple R-squared:  0.1489,    Adjusted R-squared:  0.1454
F-statistic: 43.03 on 1 and 246 DF,  p-value: 3.162e-10
```

Interpretation

ACHD – The cost to hospital decreases by 39% for a patient with ACHD as opposed to a patient without ACHD

CAD-DVD – The cost to hospital increases by 60% for a patient with CAD-DVD as opposed to a patient without CAD-DVD

CAD-TVD – The cost to hospital increases by 40% for a patient with CAD-TVD as opposed to a patient without CAD-TVD

HR-Pulse – For every 1 unit increase in HR Pulse the cost to hospital increases by 0.004 units

Cost of Implant – For every 1 INR more of cost of Impant the cost to hospital increases by 0.000006 INR

```
Call:
lm(formula = md$`Ln(Total Cost)` ~ md$ACHD + md$`CAD-DVD` + md$`CAD-TVD` +
    md$`BODY WEIGHT` + md$`HR PULSE` + md$Diabetes2 + md$hypertension2 +
    md$other + md$AMBULANCE + md$`COST OF IMPLANT` + relevel(md$AGE,
    ref = "2") + md$`IMPLANT USED (Y/N)` + md$ACHD:md$`BODY WEIGHT`)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06576 -0.14241  0.00907  0.14490  1.25041

Coefficients:
                                Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                 11.49808759  0.20937641  54.916 < 0.0000000000000002 ***
md$ACHD1                    -0.48879740  0.15522234  -3.149             0.001853 **
md$`CAD-DVD`1                0.40810337  0.08938447   4.566          0.00000808 ***
md$`CAD-TVD`1                0.34000863  0.10400582   3.269             0.001243 **
md$`BODY WEIGHT`            -0.00123674  0.00254594  -0.486             0.627587
md$`HR PULSE`                0.00406365  0.00141429   2.873             0.004439 **
md$Diabetes21                0.40758672  0.14082687   2.894             0.004163 **
md$hypertension21           -0.15452638  0.10710636  -1.443             0.150443
md$other1                   -0.14026401  0.10044371  -1.396             0.163915
md$AMBULANCE1                0.11431894  0.08005580   1.428             0.154639
md$`COST OF IMPLANT`         0.00000624  0.00000186   3.354             0.000929 ***
relevel(md$AGE, ref = "2")0 -0.00775951  0.13457874  -0.058             0.954071
relevel(md$AGE, ref = "2")1 -0.03875012  0.10204515  -0.380             0.704489
relevel(md$AGE, ref = "2")3  0.18971414  0.08258392   2.297             0.022497 *
md$`IMPLANT USED (Y/N)`Y     0.32176030  0.09884679   3.255             0.001303 **
md$ACHD1:md$`BODY WEIGHT`    0.00970086  0.00493132   1.967             0.050353 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3578 on 232 degrees of freedom
Multiple R-squared:  0.5292,    Adjusted R-squared:  0.4987
F-statistic: 17.38 on 15 and 232 DF,  p-value: < 0.00000000000000022
```

Predicting Length of Stay at ICU as it captures 68% of variation in Total cost

Step wise variable selection method was used to build a predictive model to estimate the stay of a patient in ICU for his treatment

This further helps the hospital to predict a strong variable to compute the total cost accurately

The variables choosen was same as the ones choosen to predict total cost

```
Call:
lm(formula = md$`LENGTH OF STAY - ICU` ~ md$ACHD + md$`CAD-DVD` +
    md$`CAD-TVD` + md$RHD + md$`BODY WEIGHT` + md$`HR PULSE` +
    md$Diabetes2 + md$Urea + md$AMBULANCE + md$AGE + md$`IMPLANT USED (Y/N)` +
    md$RHD:md$`BODY WEIGHT`)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8769 -1.4471 -0.4298  0.6681 25.9504

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.34793    1.42053  -0.245 0.806723
md$ACHD1                 -1.26810    0.76728  -1.653 0.099724 .
md$`CAD-DVD`1             2.99224    0.78710   3.802 0.000183 ***
md$`CAD-TVD`1             2.77159    0.90588   3.060 0.002474 **
md$RHD1                  -5.94232    2.80003  -2.122 0.034865 *
md$`BODY WEIGHT`         -0.03775    0.01881  -2.006 0.045956 *
md$`HR PULSE`             0.03827    0.01215   3.150 0.001846 **
md$Diabetes21             2.28427    1.23542   1.849 0.065717 .
md$UreaNormal            -1.25748    0.41345  -3.041 0.002622 **
md$AMBULANCE1             2.24688    0.70882   3.170 0.001728 **
md$AGE                    0.03831    0.01622   2.362 0.018993 *
md$`IMPLANT USED (Y/N)`Y  2.51420    0.69198   3.633 0.000344 ***
md$RHD1:md$`BODY WEIGHT`  0.08760    0.05641   1.553 0.121804
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.144 on 235 degrees of freedom
Multiple R-squared:  0.3668,    Adjusted R-squared:  0.3344
F-statistic: 11.34 on 12 and 235 DF,  p-value: < 0.00000000000000022
```

References

- http://www.bloodpressureuk.org/BloodPressureandyou/Thebasics/Bloodpressurechart
- https://www.mayoclinic.org/tests-procedures/blood-urea-nitrogen/about/pac-20384821
- https://www.quora.com/Why-do-we-log-variables-in-regression-model

THANKYOU !!